MACROECONOMICS

ϵ-away from complete

Marina Azzimonti, Per Krusell,

Alisdair McKay, and Toshihiko Mukoyama

with

Timo Boppart, Giancarlo Corsetti, Luca Dedola, John Hassler, Juan Carlos Hatchondo, Jonathan Heathcote, Andreas Hornstein, Pete Klenow, Simon Lloyd, Leonardo Martinez, Kurt Mitman, Conny Olovsson, Monika Piazzesi, Vincenzo Quadrini, Morten Ravn, Richard Rogerson, José-Víctor Ríos-Rull, Ayşegül Şahin, Martin Schneider, Kjetil Storesletten, and Giovanni L. Violante

Contents

Preface

Macroeconomists study a huge range of topics. Why are some countries rich and others poor? What determines the inflation rate? Why is income so unevenly distributed across people? Why do governments default on their debts? How does the economy interact with the environment? Somewhat surprisingly, it is possible to shed light on these varied questions by building on a core analytical and methodological framework. This book presents that framework and then applies it to the topics above and many others.

The book is primarily intended for first-year Ph.D. courses in macroeconomics. A key feature is that it integrates empirical evidence with theoretical analysis: data is presented throughout the book, and macroeconomic theory is consistently applied to interpret the data. This emphasis on the data-theory connection is what distinguishes the book from more traditional approaches. The book also discusses both the strengths and limitations of current macroeconomic understanding, making clear the areas in which research continues to evolve. The text has the rigor of a Ph.D. course but is presented in an accessible way, making it suitable for a wide range of students, including those in master's, or even advanced undergraduate, courses.

Parts I and II of the book present the core analytical framework and methods of analysis. These chapters were written jointly by Marina Azzimonti, Per Krusell, Alisdair McKay, and Toshihiko (Toshi) Mukoyama (the "core" authors). The chapters in Part III then build on this foundation to study key topics within the field. These chapters were written by a team of contributing authors with expertise in the areas they cover. We are immensely grateful to them for their contributions. While Part III has an array of authors, our goal was to provide an integrated text with conceptual links across chapters. We see the book as an ongoing adventure and as a great opportunity to keep engaging in scientific communication with readers around the world. We have created a website (https://phdmacrobook.org/) where you will find appendix material, data, computer code, and other resources. You are all welcome to contact us with comments and suggestions for the text and for the website.

A preface is also a place where a great number of thank yous go out to all those who have helped us make the book project come to fruition, from early-stage encouragements and patience with us for forsaking other important tasks, to research assistance as well as a broad variety of helpful comments and feedback, without any of which we would not have arrived at this point. At the risk of forgetting important people—not only economists, budding and mature, but also family and friends—because the list would be very long, we decided not to come up with such a list. Suffice it to say that we are in deep gratitude to you all.

This project is the culmination of many years of effort starting in the late 1990s at the University of Rochester, where a set of students produced lecture notes following Per

Krusell's first-year teaching in macroeconomics. With minimal editing but significant additional writing, these notes gradually became what looked like a book manuscript and it was used in a number of Ph.D. programs. However, this document was severely incomplete and became increasingly out of date. In 2021, Per had the idea of recruiting a team to produce a publishable text building on his lecture notes but going much further. The four core authors represent four different continents and, together with the contributing authors, cover a wide range of perspectives. We would like to give a special thank you to Vincenzo Quadrini who collaborated with the core team for a number of years but eventually decided to step back from this role. We are extremely grateful for his contributions early on, in addition to those in the chapter on financial frictions.

Finally, some may wonder about the giraffe. In the mid-2010s, it was clear that macroeconomics had become in severe need of an animal as a symbol of the field. The animal sought after had to have a broad overview over its habitat. An eagle? After long deliberations, the giraffe was chosen because it combines the broad perspective with standing on the ground, on firm microfoundations. Last but not least, the giraffe has the biggest heart of all animals.

Marina, Per, Alisdair, and Toshi

How to read this book

There is more material in the book than can reasonably be covered in a single year-long sequence. In our view, the chapters in Part I (Chapters 1 through 8) are essential, as they present the analytical and methodological foundations of macroeconomics, and they should be read in order. The Chapters in Part II (Chapters 9 and 10) play a supporting role—they are essential for some parts of Part III, but not others. Since most of the book is in discrete time (with the exception of Chapter 13), Chapter 9 (covering continuous time methods) is optional. Chapter 10, on the other hand, is strongly recommended, as computational methods are used extensively in subsequent chapters. The chapters in Part III do not necessarily need to be read in order, but there are cases in which one chapter builds on another. The list below gives details of how chapters depend on one another and form well-suited combinations that cover topics of interest.

- *Growth and Firms*: Chapters 9, 13, 19, 22.
- *Business Cycles*: Chapters 14, 18, 19, 20, 23.
- *Macro Labor*: Chapters 12, 14, 20.
- *Heterogeneity*: Chapters 11, 12, 21, 22.
- *Macro Finance*: Chapters 16, 19, 23.
- *Money and Monetary Policy*: Chapters 14, 17, 18.
- *Fiscal Policy*: Chapters 15, 24.
- *International Macro*: Chapters 23, 24.
- *Topics*: Chapters 15, 21, 23, 24, 25.

Part I

Foundations

Chapter 1

The subject matter

This chapter has two main goals. One is to introduce some of the central questions that macroeconomists are interested in. The second is to emphasize an important theme in the text: how our field very fundamentally is an empirical one and that measurement—the ongoing construction and improvement of data sets—and theories addressing the resulting data are intertwined and have evolved in tandem. The chapter also serves to introduce and motivate the empirical and theoretical methods that we use throughout the text.

There are several ways to define the field of macroeconomics, and there is a point to all of them, but they each also come with caveats. One definition is that the field is the study of aggregates. Here an important caveat is that macroeconomics, at least nowadays, is greatly concerned with the distribution of income, the distribution of wealth, and so on, which are not traditionally thought of as aggregates. A second definition is based on methods: macroeconomics is the quantitative study of general equilibrium. Then again, a full general-equilibrium analysis is often not necessary for analyzing many of our core issues. One example illustrating this statement is that most national economies, and even the U.S., are very dependent on the global economy, thus always making the study of a single economy partial to some extent. In addition, a more general point is that, very often, the key component of general-equilibrium analysis turns out to be the characterization of behavior—of consumers and firms—given prices, with the market-clearing mechanism playing a subordinate role. A third definition also emphasizes methods: macroeconomics is the study of the dynamics of the economy, for example by its emphasis on investment and on the role of expectations. However, as we will illustrate in this text, often a static analysis can shed fundamental light on macroeconomic issues. Less of a definition but of equal importance is the fact that macroeconomics tracks current events: the field tends to follow what is perceived as the major issues for our economies at any point in time, whatever they might be. For example, the Great Recession of 2007–2009 had its roots in financial- and housing-market fragilities, and was followed by a "jobless recovery"—a period where GDP rebounded but employment lagged significantly, keeping unemployment elevated and labor-force participation low even as output increased. Since then, these "topics" have become central to the field and attracted interest from a large number of researchers. The European sovereign-debt crisis of 2011-2012 renewed interest in sovereign default and its macroeconomic linkages—how default risk, sovereign spreads, and fiscal constraints interact with the business cycle. During the unprecedented COVID-19 pandemic, many macroeconomists put their tools to use to ex-

amine its impact on the economy and possible interventions aimed at reducing externalities from social interactions. This involved combining standard dynamic macroeconomic models with epidemiological models of virus spread. Most recently, the 2025 expansion of U.S. tariffs has brought international macroeconomics to the forefront by highlighting how trade policy interacts with exchange rates and global trade patterns via terms-of-trade movements.

The list of definitions of the field is not exhaustive and we think it is useful to use all these perspectives, in addition to the fact that a driver behind macroeconomic research is not only to understand but also to fix problems, if possible, through government policy. For now, the present chapter will be organized around the last point above: the idea is to take you through some of important events in the development of our field. This is by no means meant as an attempt to write a doctrine history; rather, the main, and really only, purpose is to illustrate the continuous development of measurement and theory to, precisely, respond to the major needs of our times. Also, we hope that the discussion will illustrate how the focus of macroeconomics keeps moving and touching base with—and, most of all, borrowing insights from—various other subfields of economics, such as labor economics, finance, microeconomic theory, and public economics.

1.1 A random walk along our macroeconomic history

We will now highlight some of the key questions in macroeconomics, along with the need for measurement and theory, and bring them up in the context of some real-world events. The focus, especially when it comes to measurement and where data is collected, is on the U.S.; for what it is worth, there is no deep motivation behind this choice, other than a practical one: most of macroeconomic frontier research has addressed U.S. data. At the same time, it is important to recognize that macroeconomic analysis typically does need to be adapted to the country under study—institutions differ, the role of foreign trade differs, the availability of data differs, and so on. In particular, perhaps, countries further from the development frontier face different macroeconomic problems and we will touch on these in Chapters 13, 23, and 24. Generally, however, we very much encourage you to make comparisons to other countries—both in terms of events and measurement—as you go through the text.

1.1.1 The Great Depression: what is going on?

About one hundred years ago, there were macroeconomists—in the sense of economists who had thought long and carefully about the performance of the economy as a whole—but these macroeconomists had virtually no systematic data to relate to; there were only scattered accounts of some production figures and some prices. What we now know as the national income and product accounts began to be developed in the 1920s and these efforts were made much more urgent by the need for systematic measures of just how badly the economy was doing in and around the Great Depression. The emerging systematic measures allowed macroeconomists to obtain partial answers to the simple question **"What is going on?"** The newly available data therefore helped John Maynard Keynes, and many others following in his footsteps, in their analysis of the macroeconomy and of how there might be ways to use government policy actively to combat recessions. In contrast to the thinking

at the time, Keynes emphasized market imperfections, in particular the sluggish adjustment of wages and prices to changes in market conditions, leaving room for a decline in the demand for consumption and investment to affect production. Simple theories connecting our now-observable main aggregates—output, consumption, and investment, private and public—were thus constructed and they are influential still today.

Constructing our aggregate data is not an easy task, however, and many of these early measurement efforts were major research achievements with important contributions from Colin Clark, Simon Kuznets, and Richard Stone. Let us now briefly describe some of the results of this work.

What do we measure?

The main purpose of national income accounts is to measure "what is *really* going on" in terms of quantities of goods and services produced. To aggregate many type of goods, we use their nominal values so the national accounts are most directly measured in nominal (current dollar) terms rather than real terms. An important question was how to compare national aggregates across time. At a general level, the answer is straightforward: we must develop a measure of the general level of prices, a price index, to separate nominal changes into quantities and prices. Once we adjust for the change in the general price level we obtain measures of real production.

Price index construction

A price index is, generally speaking, a weighted average of prices with weights corresponding to the importance of different goods as measured by the share of total expenditure spent on them. In constructing a price index, one must make some choices: the weighted average could, for example, be arithmetic or geometric and the expenditure shares could be taken from a base year or could be evolving over time.[a] These choices are often guided by microeconomic theory. For example, in Chapter 6 we will derive a price index for a consumer with a constant elasticity of substitution between goods. In that context, the price index is inversely related to the utility a consumer obtains from a given level of total expenditure.

[a] E.g., they could be straight averages of the current and last year's shares, as in the Törnqvist index.

Gross domestic product (GDP) is not a welfare measure; rather, it is a measure of how much the market economy is producing. GDP does not take many relevant sources of utility into account as they are not market-produced goods and services. Leisure, along with many environmental amenities and illegal transactions, are not counted. The difficulty in incorporating these non-market activities is that it is very hard to know how to value them. In contrast, market activities have values measured in common units (dollars).

National income and product accounts

The accounts follow double-entry methods: they measure both expenditures and income. That is, someone's expenditure is always someone else's income, thus allowing double-

checking. GDP can thus be seen either as the sum of all final expenditures (private and public consumption and investment plus net exports or as the sum of all incomes). It can also be calculated as the sum of values added by all firms. A firm's value added is its revenue from sales less its expenditure on intermediate inputs bought from other firms. The reason for not including intermediates is to avoid double counting: if one firm produces a fully equipped car, except for its exterior paint, and the second firm buys the car, paints it, and sells it, a sum of these two transactions would be close to twice the final value of the car and, thus, not be a good measure of the economy's car output. As a firm's value added measures its production, GDP is also equal to the market value of all production.

The two broad categories for income are labor income and capital income: wages and salaries and profits, interests, and dividends, respectively. Note here that profits for firms that, say, extract and sell oil, are counted in full; the resulting decline in the stock of natural resources is not deducted. For this reason, GDP is sometimes reported with a deduction for "natural-resource rents."

Capital (equipment or structures) is an intermediate good in some sense—it is bought as an investment good by a firm and used as an input in future production—but expenditures on capital are treated as final expenditures and thus counted in GDP. This observation actually suggests that when GDP is measured for longer time periods, it should perhaps treat capital differently. Capital income should, then, be a smaller share of total income, and GDP would be correspondingly smaller since it would count some of the investments—at least those that depreciate fully within the now longer period—as intermediate goods instead.[1] With this perspective, over longer time periods, GDP averages should perhaps receive less focus than average consumption.

The pandemic: what is going on?

A parallel to what occurred during the Great Depression—another **what's going on** question—took place during the COVID-19 pandemic when economies underwent massive changes overnight. Economists sought new high-frequency measures of activity in order to measure the economy in real time. These efforts often made use of transaction and commercial data gathered by companies in the normal course of business. For example, credit card companies can measure the spending patterns of their customers nearly in real time and the resulting data have made there way into macroeconomic analysis. The availability of credit card data has also proven very valuable for other reasons: to measure consumers' propensities to consume, say, when they receive transfers from the government.

1.1.2 Keeping track of long-run growth

By the mid-1950s, the economy was more or less back on its track and now there was a sufficient amount of aggregate data that it was possible to analyze its growth performance. Robert Solow's 1956 paper—the basic neoclassical growth model—and his 1957 paper on

[1] Conceptually, it should thus be the depreciation rate of capital goods that is key and that, at any frequency of sampling, determines whether it is to be regarded as capital or an intermediate good. In the 1990s, the national accounts started treating "software" as investment, as opposed to an intermediate good; the line between these two categories is sometimes quite thin.

growth accounting became the impetus for a burgeoning literature on economic growth that was both empirical and theoretical. The idea in Solow's growth-accounting paper is that one could use measures of output and inputs, along with the prices of inputs, and a basic theory of production to break down aggregate growth into the contributions of each input and, finally, a residual, which could be thought of as technical change: the *Solow residual*. It was thus due to the systematic measurement of quantities and prices that this kind of analysis could now be performed.

Which factors, then, accounted for most of U.S. growth? Solow found technical change to be of great, and direct, importance. However, in accordance with his 1956 theory, capital accumulation was an indirect result of technological change. This theory, thus, went beyond accounting and concluded that technological progress is the one (and only) fundamental reason why growth in output keeps going, and going, and going.

One reaction to the empirical finding was that whereas it was plausible that technological change is key to growth, its importance may have been overstated; after all, it was measured as a residual, and if input growth rates were underestimated, the role for technological change would not be as large. A particularly likely reason for this was that workers' skills were improving; there was an ongoing trend of increased schooling and work experience—"on-the-job learning"—was thought to contribute to worker productivity as well. In 1958, labor economist Jacob Mincer wrote his influential paper relating individual wages to years of schooling and experience and found very strong regularities in the data that have been imported and used in constructing better measures of labor input: "human capital." The so-called Mincer equation, which can be derived based on a simple opportunity cost-based theory, says in its estimated form that one more year of schooling adds a little below 10 percent to your wage. Another important measurement development in the growth-accounting literature was the notion of a firm's user cost of capital; firms buy capital and often own it until it gets scrapped, so how should one measure the "year-by-year price" of this input? Hall and Jorgenson (1969) developed an answer that was consistent with microeconomic theory and has been used ever since.

The growth-accounting literature developed virtually into an industry, where productivity performance was computed and accounted for on a disaggregated level. A natural accompanying project was to construct similar data series across countries and compare them. In 1978, Kravis, Summers, and Heston published a paper with comparable data series for 100 countries, an effort that was later continued and today takes the form of the Penn World Tables, a crucial data source for students of economic growth. In a related effort, Maddison used various sources of data to estimate GDP levels for a range of countries going back into the early nineteenth century. Harmonizing data across countries is challenging, and comparing real output too: should nominal outputs be compared in real terms by use of nominal exchange rates? Because the purchasing power of different monies vary by country, as it does within countries too, a so-called PPP adjustment gradually arose as a new standard; we discuss it and its implications in the growth chapter.

As a result of the multi-country data sets, new light could now be shed on the process of growth; in particular, **what made some countries grow so fast and others stagnate?** The endogenous-growth literature of the late 1980s and 1990s thus asked these questions, which were partly phrased as challenges to the Solow model. The endogenous nature of technological change—in particular how it is driven by incentives to innovate—as well as of

human capital accumulation came into focus, first theoretically and later empirically. Today, for example, we have access to large patent data sets that are currently under the magnifying glasses of hordes of researchers. Some of these data sets contain information on individuals so questions such as **Who becomes an inventor?** now occupy many macroeconomists.

An aspect of economic growth is **structural change**. Structural change typically refers to how some sectors shrink over time and others grow; roughly speaking, a typical path is one where countries gradually build their income—and "develop"—starting with agriculture as a dominant activity, gradually then moving into manufacturing, and finally growing the service sector. Today, agriculture only employs a percent or so of the total workforce in the U.S., whereas in the poorest countries in the world the number is 80 percent; manufacturing has furthermore been overtaken by services. Today, many macroeconomists worry about another expression of structural change: how information technology (IT) changes the workplace. In particular, **what are the implications of automationand artificial intelligence (AI)**, for macroeconomic performance, for inequality, and for the competition between firms?

Another element of structural change is women's labor-market participation, which has risen steadily and significantly in the U.S. over the whole postwar period and stands at a very high level today (not quite as high as that for men, but close). In contrast, it is much lower in many other developed countries, though it is also high in a number of countries at a lower level of development. The government sector—its size and role—can also be depicted as part of a process of structural change, as can international trade, both for a given country and for the global economy. An important ambition of macroeconomists studying medium- to long-run issues is thus to analyze the sources of these changes as well as their effects. An element they have in common is that they are slow-moving and never causes of immediate media attention, but nevertheless crucial for our economic welfare.

In sum, research on economic growth appears to have come in waves, where theory and measurement interact in very central and mutually reinforcing ways.

1.1.3 The 1970s: an oops, with stagflation, high unemployment, and more

The period up to the early 1970s was generally perceived as one of steady growth and increased prosperity. As for macroeconomic policy, the Keynesian recipes were adopted in most countries in the form of regular interventions to stabilize the economy; the term *fine tuning* was often used. Then came an "oops": the sharp recession in 1973, along with a number of severe macroeconomic problems that ended up being quite persistent. Although the cause-and-effect question is still debated, many interpret the events as a result of the oil-price hike orchestrated by OPEC in October of 1973, and the challenging era that followed was not specific to the U.S. but shared by most of the western world. Two primary difficulties involved lackluster GDP performance—along with a slowdown in productivity that at the time appeared permanent to many (the "productivity slowdown" period)—and sharply rising inequality. The effects on inequality in the labor market had different expressions in different countries: in the U.S. and the U.K., wage inequality rose sharply, while in many other EU countries unemployment rose to very high levels and stayed high for many years.

Interestingly, a new development in theory, that would turn out to have a large impact

on macroeconomics, occurred in the early 1970s, before the drastic downturns in aggregate activity occurred: the development of search models in labor economics (McCall, 1970, and Mortensen, 1972). This was to become one of the cornerstones of a theory of unemployment that later was adopted in macroeconomics as we discuss in Chapter 20. The measurement of the concept of being "unemployed" goes back in time much further, to the late 1930s; the new search theory in the 1970s thus benefited immediately from available data. However, the interest of macroeconomists in the topic rose sharply during the productivity slowdown period and generated further data needs. Today, a central part of our analyses of unemployment includes detailed data both on individuals and on firms, most of it in survey form. Moreover, keeping track of inequality trends, such as the average wage gap between skilled and unskilled labor that took off beginning in the second half of the 1970s, has become a central activity for macroeconomists.

Labor-market data

Much of our information on the labor market comes from surveys. Total labor income is available from NIPA but how it breaks down into employment, hours worked, and wages/salaries for different workers is all based on how individuals answer questions in questionnaires. A key source is the Current Population Survey, conducted monthly by the BLS in collaboration with the Census Bureau. The CPS has a (limited) panel feature, i.e., it interviews the same people at more than one point in time. The CPS also measures unemployment, i.e., it asks people if they are not working and looking actively for a job, thus in line with search theory. An important source of data on individual labor-market outcomes is the Panel Study of Income Dynamics (PSID), conducted by the University of Michigan, started in 1968: it follows individuals over a significant amount of time and has data on a number of individual variables, all self-reported. The breakdown of labor earnings into hours worked and a wage per hour in these data sets is based on individual reporting on how many hours they work. Other panel data sets on individuals include the National Longitudinal Survey (NLS, conducted by the BLS) and Survey of Income and Program Participation (SIPP, conducted by the Census Bureau).[a] Firms are also surveyed and report work hours for their employees (e.g., in the Annual Survey of Manufactures from the Census Bureau), but then the same workers may have multiple jobs so measures of how much an individual works in total still rely on asking the individual. Around the turn of the millenium, and in part as a result of the new developments in the field of macro labor, where search frictions for firms and workers are in focus, BLS also started to collect micro data, published from 2002 and on: the Job Openings and Labor Turnover Survey (JOLTS). This data set has since become invaluable in the evaluation of further developments of our theories.[b]

How individuals spend their total amount of time is also measured. Since 2003, the BLS has produced the American Time Use Study (ATUS), a survey documenting daily activities in great detail, including how much time is leisure versus various forms of "work at home." Using ATUS, it is also possible to find out not only whether people searched for jobs, but how much time they spent on this activity.

[a] A longitudinal survey is a panel, i.e., a study that follows the same individuals over time.

[b]Today, data from labor markets—e.g., recent trends, particular skills in demand—are also provided commercially, often with real-time information from the internet, and used by human resource departments and head-hunting agencies.

Governments attempted to stabilize the fall in GDP, in particular with expansionary monetary policy, i.e., by cutting interest rates. However, there was very limited success; rather, the 1970s was also a period of unusually high inflation (with annual rates in the 10–20 percent range in many countries, and even higher for some). The combination of stagnation and inflation was dubbed *stagflation*. The Keynesian paradigm came under increasing scrutiny and a number of economists focused on weaknesses in the Keynesian theory itself. A particularly powerful point was the "Lucas critique," which explained how reduced-form relationships between aggregates—a cornerstone in the applied Keynesian apparatus—could break down if policy changed. The Phillips curve—the negative relationship between the inflation rate and the unemployment rate—was a particular case in point: in a paper that actually predated the oil crisis and stagflation, Lucas (1972) advanced a theory showing how attempts to exploit this seeming trade-off with monetary policy would make the relationship itself break down. When the relationship did break down, Lucas's sharp critique gained added force and, in hindsight, marked a clear break in the development of macroeconomics.

Curiously, the 2021–2022 period echoed the 1970s: a sharp rise in headline inflation, amplified by an energy price shock linked to the Russia-Ukraine war. This prompted fears of a new stagflation period. However, major Central Banks around the world raised interest rates sharply in 2022–2023. By late 2023 and into 2024–2025, inflation fell back as energy costs eased and supply bottlenecks decompressed, while labor markets stayed relatively tight. As a result, the feared stagflation did not materialize. Undoubtedly, macroeconomists are better equipped today in confronting these kinds of episodes, but every challenge seems to have its unique properties.

1.1.4 Kydland and Prescott: a way forward

Lucas's critique was not just destructive, in pointing to weaknesses in the Keynesian theory approach, but he suggested, at least conceptually, how an alternative framework could be built up. The idea was to build explicitly on microeconomic theory, and with Kydland and Prescott's 1982 paper (see Kydland and Prescott, 1982a) it became clear just how to do this in a way that also allowed systematic comparison with data: they offered *quantitative theory*. This paper suggested basing the microeconomics on empirical studies in applied fields, such as labor economics and consumption studies, not just in terms of the structure but also by importing parameter values from the empirical microeconomic literatures. Kydland and Prescott's paper, which led to an explosion in macroeconomic studies, also added another important aspect of measurement: that often, data needs to be detrended, or "filtered," in order to be ready for analysis, in their case in examining the sources of business cycles.

Filtering

When macroeconomic models are built and compared to data, the data are almost al-

ways filtered first. To understand what filtering means, you need to see macroeconomic models as dynamic systems, i.e., as some form of vector difference equations, that contain random variables. Thus, macroeconomic models in fact define a *stochastic process*. Such a process could thus be simulated by the researcher and, in principle, compared to data. However, the idea is rarely that the theory is constructed to explain everything. Kydland and Prescott, for example, were interested in recessions and booms, which are movements upward and downward in macroeconomic aggregates around some overall trend, but this trend was not the subject of their study. In order to compare theory and data, most researchers therefore use filters to extract the aspect of the data they are interested in analyzing. To do this, theory comes in handy: stochastic processes can, quite generally, be thought of as sums of sub-processes, each one with a different frequency, i.e., periodicity. For business cycles, one thinks of periodicities of between 1.5 and 8 years perhaps, and so-called band-pass filters offer can be used to take any data series and remove any frequency outside a specified range. In contrast, studies of medium- to long-term movements in variables require removing high frequencies and retaining low to medium frequencies.[a] In financial economics, when day-to-day or minute-to-minute changes in stock prices are analyzed, all but the very high frequencies are removed before the data can be analyzed.

[a] Kydland and Prescott (1982a) used a specific filter: the so-called Hodrick-Prescott filter, which is an intuitive way of extracting data. It has very wide-spread use in macroeconomics.

Kydland and Prescott's theory of the business cycle was quite stylized and stripped down—among other things, it was phrased entirely in real terms and had no role for monetary policy—and the literature that followed enriched their framework in a multitude of directions. A key point here is that the first wave of models had perfectly working markets; later, a number of frictions were added and today, virtually no macroeconomic model that is used in practice is free of market imperfections. A key friction that was added was price stickiness: firms setting dollar prices of products face costs in doing so, and therefore only adjust prices infrequently. This makes monetary policy have direct effects on the economy, something it might otherwise not have. Thus, the "New Keynesian" framework was built up, where monetary policy is in focus. Again, the new theory led to measurement efforts. In particular, studies such as Bils and Klenow (2004a) looked at the survey data available underlying the CPI and recorded the frequency of price adjustments; their work allowed researchers to parameterize the microeconomic structure for adjustment costs assumed in the models.

1.1.5 Different waves of macroeconometrics

The comparison between models and data has also undergone waves. As with much of economics, it is challenging to discern causal relationships in the historical data given to us. As more and more data have become available, however, more and more thinking has been devoted to the development of different statistical methods for doing this. A central question has thus been the purely methodological one of how historical macroeconomic data can be used to make inferences; for this reason, we devote significant space in this text to the methods used today (see, especially, Chapter 8).

Macroeconomic models are all simplifications of a highly complex system and therefore there is little point in "testing a model" by assessing whether it can be the true data generating process. As the saying goes, all models are wrong but some can be useful. A useful model allows us to answer an important question in a convincing way and the model must be consistent with the relevant data we can observe in order to be convincing. Of course, if the answer to our question can be directly observed then there is no point using a model. So a useful model allows us to bridge the gap from the data we observe to the questions we want to answer. Knowing which data are important to match in order for the answer to be convincing is often argued to be an art. But art, too, can be taught.

The empirical implementation of Keynesian theory involved estimating large systems of (usually linear) relationships, often with ad-hoc specifications of short-run dynamics, i.e., with lags of variables added so as to provide a better fit. Sometimes instrumental variables were used, but that was more uncommon. The critique that came in the 1970s forced macroeconomists back to the drawing board. One approach was to estimate the new, now microeconomics-based, structural models that rapidly developed using a classical statistical methods. A literature using maximum-likelihood and Bayesian techniques for estimation was developed; a related development involved use of the generalized method of moments, which could be applied to a subset of the model's equations.

Another theory-based path, labeled calibration, was the method favored by Kydland and Prescott in their work. The calibration approach is very common in current macroeconomic research and can be used to derive quantitative conclusions from a theoretical model. For example, in their work on business cycles, Kydland and Prescott wanted to know to what extent movements in technology could generate fluctuations in aggregates that resemble those in the data. The spirit of calibration is to select the model's parameter values based on other moments of the data than those in focus in the study. For example, Kydland and Prescott based their parameter choices on two kinds of data: (i) micro data, e.g., for people's attitudes toward risk and intertemporal substitution; and (ii) long-run facts (that is, low-frequency data not in focus for their high-frequency interests). Once the model parameters have been selected, one can then derive the model's predictions for the phenomenon of interest. For example, Kydland and Prescott calculated the variances and correlations of aggregate variables to assess whether the model could generate business cycles that resembled those observed in the data. The sentiment of "all models are wrong" may explain the wide use of calibration within macroeconomics. Calibration does offer discipline in the sense described above—parameters are not to be chosen to match the moments the researcher wants to explain—but, at the same time, does not lend itself to hypothesis testing.[2] Relatedly, as all models are wrong we are more interested in the broad patterns they predict: can the model at all account for the phenomenon under study, or are the magnitudes severely off? If the model can generate patterns similar to those in the data, it is typically judged "potentially useful" and elaborated on further, possibly adding detail and examining auxiliary implications. This is how "technology shocks" entered our vocabulary and are, still today, considered relevant for (but far from alone in) explaining business cycles.

[2]The lack of hypothesis testing is shared with Bayesian analysis. Of course, in Bayesian analysis, the parameter selection is informed by the data under study; in calibration it is not: all the weight is on the prior.

A much less structural approach was proposed in Sims (1980a): vector autoregressions (VARs). Sims's focus was much more on the identification of causal effects in aggregate data, and the core of his methods involved ways to observe plausibly exogenous shocks, such as an unexpected increase in the Fed funds rate, and then trace out the effects of the shocks on macroeconomic variables, including the effects on the subsequent movements in the rate. In its simplest form, a VAR is a linear system of variables including lags, thus containing both intra- and intertemporal relationships, and a shock to each variable at each point in time. A literature also evolved that took structural models and derived their linearized VAR approximations, which could then be compared to the estimated VARs as well as offer some structural interpretations of some the coefficients in the VAR. VAR analysis is a very common tool in macroeconomics.

Another approach to identifying causal effects is to use natural experiments. Increasingly, macroeconomists use information from natural experiments to identify the strength of causal relationships at the microeconomic level and use these moments as targets when calibrating a macroeconomic model. An important example is measuring the marginal propensity to consume out of wealth. In 2001, the U.S. government paid tax rebates to most households and randomly gave some households their payments sooner than others. Johnson, Parker, and Souleles (2006) used the random timing of the payments to measure how strongly consumption spending responds to additional income. This type of information is now an important calibration target for many macroeconomic studies as the marginal propensity to consume is important for understanding the effects of certain government policies. In this case, the natural experiment occurs at the level of the household as some households are paid earlier than others. Natural experiments at the level of an entire economy are more rare but there are some examples. One type of application is exemplified by Acemoglu, Johnson, and Robinson (2001), who used excerpts of historical records from the colonial era to make causal statements about the effects of institutions on long-run economic growth and well-being. Another example is the "narrative" approach to the evaluation of monetary policy, where minutes from Federal Reserve meetings are analyzed to identify exogenous events affecting Federal Reserve policy (Romer and Romer, 1989). The Romer and Romer study is an example of another phenomenon, which is to use text analysis (e.g., words used in media) in macroeconomic contexts; the wave of big-data tools has thus also entered our field.

1.1.6 Models: intuition vs. quantitative use

Kydland and Prescott's work was a game changer also in how macroeconomists approach model building, which previously had been largely oriented toward building an intuitive understanding of mechanisms—such as Lucas's 1972 Phillips-curve paper. Of course, there are a huge number of different mechanisms at play, so how does a macroeconomist oriented toward giving policy advice choose which mechanism(s) to focus on? Kydland and Prescott's answer was to move away from building models aimed at intuition in favor of larger models that could be parameterized and calibrated to deliver quantitative output. This quantitative output would then, within a single model, aggregate across the many mechanisms inherent in the model. The analysis of larger, nonlinear, models is much harder and, with sufficient complexity, impossible to undertake with "pencil and paper." Thus a new sub-field of macroeconomics developed rapidly: that focusing on solving dynamic models with numerical

techniques. Nowadays, the most common approach in applied macroeconomics is arguably to formulate rich models allowing several mechanisms believed to play a role, solve the models numerically, and then simulate them to study model output and compare different mechanisms quantitatively. This approach has not replaced the need to formulate much smaller models to build intuition, but the aim is to ultimately be equipped not only with an intuitive understanding of what goes into building the macroeconomic equivalent of a bridge but also with quantitative assessments that allow us to cross the bridge without fear.

1.1.7 Macroeconomics and inequality

As already mentioned, the late 1970s saw sharply rising wage inequality, a phenomenon that has continued, though with different intensity, and hit different groups differently, during different decades. **What explains these developments?** Technological change, increased exposure to trade, or changes in unionization? Many macroeconomists have turned their attention to this question. They have developed theory and examined how different theories match the data, thus making the overlap with labor economics a particularly vibrant one. Here, data on individuals has been a key input. Again, most of this data is taken from surveys, but it is increasingly common for researchers to make use of administrative data, i.e., data on the whole population.

Administrative data

The basis for taxation of individuals and firms is reports on incomes and transfers (including bequests). Here, employers report wages and salaries paid out for all taxable individuals, and individuals complement these data. For example, the self-employed report their own earnings. Similarly, firms and government agencies report transfers made, such as social security payments, and these are based on earnings. Tax authorities also have records of capital income for all individuals—dividends, interest, and capital gains—but the specific assets are not recorded. Relatedly, there is no administrative data on wealth in the U.S., since it is not taxed, but some countries do employ wealth taxes and, therefore, administrative data on wealth can, in principle, be accessed there. Thus, all the underlying data is registered and potentially a source for researchers to use. Access is restrictive, but can be granted. The Internal Revenue Service (IRS) and the Social Security Administration (SSA), for example, have been employed to provide detailed account of the distribution of incomes for the full population.

Of course, tax records and other administrative data are not freely available and, depending on the particular data set, may be more or less difficult to obtain permission to use. In all cases, the data the researcher is given is made anonymous: individuals' identities are never revealed. A particularly interesting possibility for researchers is to link different data sets (whether a administrative data set or not) but this is rarely allowed.[a]

[a]The reason is the risk of inadvertently compromising anonymity. There are exceptions. For example, in Sweden, local researchers can cross-link administrative data sets; applications to do this require careful descriptions of the purpose of the study, the methods employed, and involve various ethical considerations.

In one strand of the literature, the focus has been on data sets that contain detailed information both on firms and on their employees, hence shedding light on what kinds of firms "match" with what kinds of workers and how wages are then set and vary over time. Although the share of total income paid to labor has been remarkably stable during the post-war period in the U.S., it has had a **recent trend downward** over the last decades, a trend that can be observed also in many other countries. Hence, macroeconomic researchers are examining various hypotheses for this phenomenon, such as structural change and technological change, possibly along with changes in the degree of competition are being examined. To this end, access to data on firms is critical, and the overlap with the field of industrial organization is evident.

Firm data

Various firm-level (and establishment-level) data sets are provided by both governmental institutions and commercial vendors.[a] In the U.S., the Census Bureau and the BLS provide data from administrative sources and from surveys. We can thus obtain information about (among other things) entry, exit, and employment dynamics of firms and establishments at annual and quarterly frequencies. Although micro-level data are often confidential, many useful summary statistics are publicly available.

Some of the data sets include information about inputs and outputs on the firm level. Among other things, these can be of use for estimating firms' marginal costs, which are never directly observable; their movements over time are important for understanding macroeconomic phenomena. Relatedly, innovative activity can be measured based on firms' R&D expenditures, and the patents generated can be accessed from the patent office. As patents build on earlier inventions, the number of times a patent is subsequently referenced by later patents gives a measure of its impact.

[a] Commonly used commercial vendors include Compustat, Orbis, and NETS.

Over the last two decades, we have also observed a sharp increase in macroeconomists' interests in **wealth inequality**. Just like during the Great Depression, a common perception has been that of increasing gaps between "rich" and "poor," along with various forms of polarization, but what does the data—to the extent we even have it—really say about wealth inequality? There are (at least) two reasons for macroeconomists to care about this question, and about the underlying trends driving wealth inequality to change over time. One is an intrinsic interest in inequality as a key aggregate phenomenon: the view that it is an undesirable feature in society and should be taken into account even if it is in conflict with other goals. An additional intrinsic reason is political stability: as expressed in Piketty (2014), one may worry that democracy is threatened if inequalities rise above certain levels. As a final example, one dimension of inequality of relevance in macroeconomics is that between women and men and across racial and ethnic groups; we are now seeing an increasing number of contributions documenting and analyzing, in particular, how the relative wages and the relative hours worked across groups have evolved over time.

A second reason for macroeconomists to care about inequality is that it captures heterogeneity that is important to take into account when examining the workings of the macroe-

conomy. When, for example, a tax rebate is implemented with the purpose of stimulating consumer spending, we have reasons to think that cash-constrained, poorer households would spend a large fraction of the rebate whereas richer households will save most, if not all, of it. Thus, distributional data on wealth appears as a determinant of the efficacy of many policy interventions. The development of so-called heterogeneous-agent models, which began in the 1990s and has generated a very large literature, is a response to both these reasons to keep track of, and understand, wealth inequality.

Measuring individual wealth

In the U.S., since wealth is not taxed, there is no direct administrative data on it.[a] The Survey of Consumer Finances (SCF), also conducted by the Federal Reserve Board, is available every three years since 1983 and has data on individual assets; it is a key source of information about the wealth distribution. Unlike some of the surveys mentioned above (such as the ASM), this survey is voluntary, but efforts are made to make it representative. The IRS administrative data has capital income, so it is possible to estimate wealth by observing the annual income it generates—if one is willing to assume a rate of return on the wealth. This is called the *capitalization* method. More recently, measures of wealth inequality across countries have been put together by the World Inequality Lab. Wealth series are constructed by combining national accounts and sectoral balance sheets with household surveys, fiscal/tax data, and billionaire "rich lists" to estimate market-valued net wealth (assets minus liabilities), and particularly at the top of the distribution (See https://wid.world/).

[a] A small set of countries have taxed wealth over various periods in time and therefore have administrative data on it.

As it turns out, the various different sources used by different researchers do indicate a rather significant increase in wealth inequality in the U.S. beginning in the late 1970s. Similar trends have also been documented in a number of other countries. In sum, macroeconomics today, including at the level of policy making, is concerned with a much broader view of inequality than in the past.

1.1.8 Taxes and government activities

Many western countries, including the U.S., have experienced slow, long-run increases in the role of government, both when it comes to its total share of GDP and employment and in terms of transfers, such as social security and welfare systems more generally. Thus, the **nature, determinants, and effects of taxation** have become a central theme for macroeconomists. In the U.S., marginal tax rates were increasing and peaked shortly before Ronald Reagan took office and thereafter the degree of progressivity was lowered significantly, and has stayed at a historically low level since. The taxation of corporate profits has also changed over time. How have these changes affected hours worked, economic activity, and inequality? These questions preoccupy many macroeconomists.

Aside from these changes over time, an overall key question, aside from the degree of

progressivity of the tax code, has centered around the choice between different tax bases: taxes on capital income, taxes on labor income, corporate taxes, indirect taxes (such as sales taxes), seignorage, property taxes, and so on. That is, macroeconomics and public finance intersect in important ways. At the same time, the expansion of government activity has increasingly been financed through public debt, prompting a large literature on how debt interacts with the macroeconomy, such as potential crowding out of private investment and capital accumulation, concerns about debt sustainability and fiscal limits, and, in some settings, the risk of sovereign stress and default.

When it comes to the efficiency features of different tax rules, there is also an important overlap with economic theory. For example, it may be tempting to use tax and transfer schemes to fill in where private insurance markets appear to be missing—thus improving the economic situation of those experiencing unexpected adverse events—so macroeconomic policy making may need to address moral hazard and adverse selection. Another challenge, which has been a major issue in macroeconomic research at least since the 1970s, is the fundamental inability of governments to commit to its future policy choices. For example, basic public finance theory says that it is efficient to levy taxes on already installed capital—since it is not distorting any choices—but, if firms/investors know in advance that their capital will be taxed in the future, current investment is distorted. Thus, the government would like to say that it will not tax capital in the future, but, ex post, change its mind. This is known as a *time inconsistency* problem. The tensions inherent in differences between a government's optimal plans and their ex-post temptations to change them has generated another vibrant area of research with significant components of economic theory. More broadly, the interactions between macroeconomics and politics is another relatively recent sub-field of macroeconomics where significant research is carried out.

1.1.9 The Great Recession: another oops

In 1979, in an effort to end the high-inflation era, the new Fed chairman Paul Volcker announced, and implemented, a period of very tight monetary policy. Most macroeconomists attribute the ensuing recessions in the early 1980s to this change in monetary policy stance. Inflation did come down, and it did so also in most other western countries; the stagflation period had been a worldwide phenomenon. In Europe, steps were gradually taken toward tighter monetary policies as well and a currency union was eventually created: in 1999, ten national currencies ceased to exist and the euro took their place. This was a period of financial integration not only among European countries, but also between developed and emerging economies. Restrictions to the movement of financial assets, goods, and services were loosened, triggering a "globalization" process that generated great interdependence among economies. From the mid-1980s and for over two decades, until 2008, the U.S. economy experienced recessions but they were minor and the overall aggregate performance was viewed to be very satisfactory. In 2002, macroeconometricians James Stock and Mark Watson dubbed this era the Great Moderation: a period of time when the macroeconomic aggregates displayed healthy growth and very low volatility. By some, the Great Moderation was attributed to the new and transparent policies followed by the independent Fed. Researchers also advanced other hypotheses that were more structural, such as changes in the nature of technological change. Furthermore, the stabilization frameworks used at many

central banks now relied on the New-Keynesian macroeconomic model: a setting based on microfoundations and including a number of frictions, most importantly sticky prices and sticky wages. A prescription from these models was for the central bank to systematically counteract macroeconomic shocks so as to lower volatility and improve our welfare.

Whatever may have caused the Great Moderation, there is no doubt that few expected the events that surprised the world in 2007: a severe economic downturn that was worldwide as well. The recession was nowhere near as deep as the Great Depression, but it was nevertheless a very problematic period: unemployment rose sharply and only fell back very slowly in a manner that was uncharacteristic, compared at the very least to recent experience. The crisis immediately hit Europe as well, and in 2009 a multi-year debt crisis, sometimes labeled the eurozone crisis, was set off. A number of European countries thus suffered from high national debt levels and difficulties in rolling over their debt; this period was one of significant uncertainty. The uncertainty partly involved what paths government policy would take, and there was ample speculation that some countries would leave the currency union. In the end, they did not, but the crisis was long and painful, as were the macroeconomic and political debates about whether debts should be forgiven or not. The crisis highlighted the potential problems of a globalized economy, triggering some countries to 'close down.' The exit of Great Britain from the European Union, labeled "Brexit," was the most notable example. An important role in combating the crises was played by central banks, but now with methods very different than those pursued during the previous decades.

A period of time of intense research followed. **What were the deep causes of the Great Recession?** How could it have been avoided? Were our main theories flawed? This research is still ongoing so it is hard to draw definite conclusions, but there is consensus that a combination of excessive risk-taking in housing markets—arguably rooted both in private and government decisions—and severe frictions in financial markets together slowly sowed the seeds of the downturn. Indeed, the term Global Financial Crisis is equally often used to refer to these developments.

As a result of the experiences during this period, massive research has gone into studying the workings of financial markets and financial institutions and how government regulation affects their performance. It reminded us how asset markets, debt buildups, and (excessive?) risk-taking can be intricately intertwined with the workings of the macroeconomy. The downward trend in the real interest rate is a related phenomenon. It is suggested to be connected to more severe asset-price fluctuations, including the formation of price bubbles; here the emergence of cryptocurrency is perhaps particularly noteworthy.

Financial data

One of the key questions in this part of macroeconomics is *financial stability*. In particular, if key financial actors have strong interdependencies in their asset portfolios and liability structures, then domino effects can occur, whereby relatively minor shocks can have severe aggregate consequences. As an example, the house-price decline and the resulting defaults on mortgages in the U.S. during the Great Recession were significant, but their quantitative magnitudes were quite small compared to many other asset-price movements throughout history that barely even generated recessions at all. The reason

why the small shock had major consequences was to be found in how mortgage liabilities had been packaged and distributed among key financial actors, all of which was quite opaque not just to policymakers but to market participants as well. Hence, the need for data in this area is, and was, great, and a major challenge is that high-frequency data on portfolios is proprietary information and, when it is available, can be hard to interpret.

In its Flow of Funds section, the Federal Reserve Board produces the quarterly Financial Accounts of the United States, a comprehensive set of accounts that includes detail on the assets and liabilities of households, businesses, governments, and financial institutions. These are aggregate data, allowing us to track, among other things, trends in indebtedness—which was key in the analysis of the Great Recession, but also not quite sufficient for detecting interdependencies among financial institutions. There are also data on individual households and firms; the SCF is discussed above.[a]

Asset prices and returns for publicly traded firms are of course available from numerous sources. Assessments of values for non-traded firms are much harder to come by, and even many large firms are not publicly traded. In the past, a typical path from the birth of a firm to an established, large company involved a mix of individually provided funds, bank loans and possibly bond issues, with an eventual public offering and public trading. Today, the path toward public trading of the firm's equity often takes longer and goes via private equity financing, e.g., via venture capital companies.

How private individuals make portfolio decisions is another important input into how the macroeconomy works, but raw data on this is much harder to come by; as discussed above, data on wealth management is typically not available except in limited surveys.

[a] An often-used financial database for firms is Compustat, a commercially provided service containing information about publicly traded firms (also outside the U.S.). An even larger database is Orbis, which also contains non-traded firms, including smaller businesses.

Though it seems clear by now that the basic macroeconomic framework was not abandoned as a result of the Great Recession, it is equally clear that it has been changed and enriched in the direction of including financial frictions that play a prominent role. Macroeconomists are, perhaps painfully so, aware that the next recession will rarely have the same characteristics as the most recent one, and as a result their theories grow richer and more complex, rather than themselves undergoing cyclical fluctuations.[3]

1.1.10 Global interactions

The Great Recession was a **global** financial crisis: stresses originating in U.S. housing and banking spread to other countries through funding markets, cross-border bank exposures, and the collapse in trade and trade credit, synchronizing recessions worldwide. This experience underscored the importance of studying open-economy models, where shocks in one country (such as TFP disturbances) can propagate to other countries through shifts in capital flows and the resulting movements in relative prices (e.g., terms of trade and real exchange rates), with the strength of spillovers shaped by the degrees of international risk

[3] Voices were certainly raised suggesting that a return to classic Keynesian theory was called for.

sharing and exposure in balance sheets. The European sovereign-debt crisis in 2011-12 was triggered, in part, by the Great Recession, which exposed fragilities in the euro-area financial system. Concerns about debt sustainability pushed spreads sharply higher in countries such as Greece and Cyprus, which eventually had to restructure their debt. Fears of sovereign default extended to other European countries, such as Ireland, Portugal, and Spain who also saw their spreads raise significantly. This episode brought sovereign-default models— originally developed to study emerging countries during the sovereign debt crises in Latin America during the 1980s— back at the center stage. Macroeconomists have since spent significant effort in understanding the international transmission of business cycles and how borrowing, default risk, and default interact with fiscal policy and the business cycle. The post-pandemic rise in public debt ratios across developed and emerging countries has renewed (and keeps salient) the importance of these topics.

1.1.11 Climate change and energy economics

The intersection between macroeconomics and environmental economics was close to empty until it became clear toward the end of the 20th century that **climate change**, at least in important part caused by human emission of carbon dioxide into the environment (primarily by burning fossil fuels like oil, natural gas, and coal), was a potentially critical threat to our welfare. Increasingly, many economists have become engaged in climate research and contributed importantly to our understanding of how climate change interacts with economics. Macroeconomists have helped in damage measurement by studying aggregates and how they react to weather as well as climate. But macroeconomists have also contributed by constructing global economic models aimed at *integrated assessment*: examining how different economic policies would influence the world's market economies and jointly determine climate and economic outcomes, hence providing advice for policymakers. This endeavor has led to an upsurge in quantitatively oriented modeling of global interactions between the macroeconomy and the climate.

The focus on fossil fuels also directs the attention of macroeconomists to natural-resource and energy economics, which attracted attention already in the 1970s as the oil shocks hit. The questions now are similar to what they used to be, and they have been underscored by recent events such as the war in Ukraine. In short: **what is the nature of energy supply and how does the economy react to shortages in it; and what are the potentials for technological change in this area?** The climate-energy area, along with communication of the research results to policymakers, can turn out to be of particularly high value, given the large global interest in it. Climate science and research on energy technology are obviously crucial inputs, but policy is also needed to steer individuals and markets in useful directions. How this is best done is a matter of understanding economic decisions and market interactions: it is a question for economists.

1.1.12 Where do we stand?

Clearly, our economies are constantly evolving as a result of a number of societal changes, including technological developments and policy reforms. With these changes, we have indicated how macroeconomics has changed course, sometimes abruptly and sometimes merely

by expanding on existing frameworks. We currently have a body of knowledge that allows for a more nuanced understanding of macroeconomic events and policies, and the macroeconomic models used in practice are accordingly much richer than in the past.

Do we lack sufficient self-criticism, however? One regularly hears arguments that macroeconomics never really admits that it is wrong, nor that it recognizes that it needs to change course. In concrete terms, can we really say that we are in a better position today to meet the next major macroeconomic challenge? Our perception, first, is that macroeconomists do admit mistakes, as indicated by the number changes in our thinking that have been described above. On some occasions, our theories simply have not incorporated all relevant features—this would be "our" mistakes—and as a result we have tried to build these features in as swiftly as possible. On other occasions, though, we have simply been surprised by events that did not have economic origins and yet necessarily generated economic downturns that, once we were informed of the "shock," we could understand with existing models.

Our own firm belief is in fact instead that, guided by research, macroeconomic policy has been increasingly successful over time. Not all recessions are alike, but they are not all distinct either, and lessons from one will typically be useful in the future too. In particular, we count the responses to the Great Recession and to the 2019-20 pandemic recession, where fiscal and monetary policy were used actively, as having benefited in major ways from macroeconomic research. At the same time, of course, not all governments acted alike and there always remain differences in views on policy, especially since strong views on economic policy tend to be linked to strong political views.

Thus, as different shocks hit our economies, we do not cycle back and forth between models, each model emphasizing one type of shock and how to respond to it. Rather, we strive to combine insights as they arrive and try to isolate how they add to, rather than erase, our previous understanding of the economy. For illustration, consider Keynes's core insights: they were entirely new and crucial for building up an understanding of how stabilization policy could be usefully conducted, and although the 1970s saw a definite break in the foundational elements underlying macroeconomic models and a temporary return to very stylized models, the Keynesian insights have since been added back into the newer models. These models are simply more sophisticated today than before and are more clear on the circumstances under which Keynes's insights can be applied. The models keep developing; for example, the reliance on rational expectations may develop, as we obtain more data on how forecasts are actually made. Macroeconomists appear to be in agreement that frictions in financial markets can be critical and even central in explaining some recessions, but how to identify the key frictions and prevent them from playing out is still an open question; clearly, a prohibition of all loans would by definition have prevented the mortgage market from causing problems, but very clearly such restrictions are highly undesirable. Thus, research on this issue today is trying to identify how we can reach a balance between stability and business-as-usual market efficiency. Many challenges remain in macroeconomic research but we are nevertheless optimistic that there will be fewer and fewer oopses in the future.

Finally, the construction of rich and complex models is of course not an end in itself and especially for communication and in teaching—even at the graduate level—it is important to simplify and abstract from many of the complexities. Therefore, IS-LM models can still be useful in building an understanding of how the macroeconomy works, as can real business cycle models and simple models of debt crises. But beliefs that either of these models is

sufficient has been proven wrong many times over.

1.2 Looking ahead

This introductory chapter has sampled a number of important topics addressed by macroeconomists over the course of the last century. Many topics have been left largely without comments, such as trade liberalization and immigration. This is not to suggest that they are any less important: they remain very active research areas, some overlapping with further sub-disciplines of economics such as international trade—and the main focus here has instead been to illustrate the large variety of topics and methods.

The development of rich and complex models means that macroeconomics is not becoming easier. This textbook is a living proof of this statement: although it makes a major effort to mix data and historical episodes with theory, it does require new students of this area to make serious investments in methods. In particular, the switch toward microeconomics-based theory, with an accompanying aim to match the main historical facts—*quantitative theory*—motivates many of the early chapters. The belief on which this text is based is that this material is here to stay. One possibility is that the theory will be supplemented with elements of behavioral economics, but there is little consensus yet on which features are key. This regards both how we view the choices made by consumers and firms and how they form expectations. For now, full rational optimization and expectations are viewed to be a very reasonable starting point, and the insights from investments in these methods will also highly likely be relevant as behavioral elements may be added. Thus, we try to view the methods parts as fun, because they are, though they are never there for their own sake: their only aim really is to help us understand the macroeconomy. This is the nature of macroeconomics; it is challenging, but it is engaging.

The overall text has two basic parts. The first one introduces the main methods and framework used to make sense of the core facts of macroeconomics. This part is core material and should be read in advance of the rest of the text. The second part has applications. The set of applications perhaps contains slightly more material than is covered in a typical first-year PhD course, but only slightly; the recommendation is nevertheless to read the chapters in order. In the first part, Chapter 2 is of particular importance since it is a lead-in to the rest of the text. In particular, it dives into the macroeconomic data and, bit by bit, introduces Solow's way to make sense of this data: macroeconomic aggregates are generated by the neoclassical growth framework. This framework is the core setting used in macroeconomics. The framework is not an arbitrary one: as the chapter explains in some detail, it is instead precisely motivated by a need to square our theory with some striking, long-run facts that are very hard, if not impossible, to explain without this theory. The chapter also moves beyond Solow's treatment by arguing that some parameters he treated as exogenous constants—saving rates and hours worked—are better described as conscious choices of households in a market environment. The chapter concludes with a preview of the remainder of the text, emphasizing that virtually every chapter thereafter, dealing with the main, applied topics in macro (growth, business cycles, asset prices, labor markets, etc.), build directly on the market version of the neoclassical growth model.

Chapter 2

A framework for macroeconomics

The purpose of this chapter is threefold. First, we go through the main macroeconomic time series, focusing primarily on their historical properties. We mostly use U.S. data and thus encourage the reader to examine the corresponding graphs for other countries.

The second, and key, purpose of the chapter is to gradually introduce, rather heuristically and with the formal details postponed until the ensuing chapters, the basic framework—the macroeconomic model—that will constitute the core tool in the textbook. Thus, each graph will be interpreted from the perspective of the proposed framework. An underlying assertion is that it is hard, if not impossible, to account for the data except with the kind of framework we use. This framework goes back to Solow's growth model and then builds in conscious choices, such as firms' choices of inputs, consumers' choices for saving and hours worked, later on the purposeful development of human capital and technology, and so on. The emphasis on explicit choices necessitates a microeconomic approach, not just in terms of theory but also in terms of the data we will look at; nowadays, much macroeconomic research directly studies cross-sectional data (for households, firms, etc.). There is no presumption that markets work perfectly. Instead, much of the analysis, especially when it comes to looking at macroeconomic policy, centers around pinpointing specific weaknesses in the functioning of markets. Finally, a key feature of the core framework is that it is *quantitative*: the aim is to formulate a model that can account for the magnitudes of macroeconomic phenomena and not just their qualitatively features.

The third purpose of the chapter is to be a stepping-stone into the rest of the text. Thus, this chapter will offer a brief description of the topics studied in later chapters.

2.1 The facts and interpretations: real aggregates

In this section we document some basic facts relevant for macroeconomic analysis. The facts are presented in a stylized manner; for example, the unemployment series will be described as "stationary" and this term should not be interpreted in a statistical sense but rather as a series that does not have a marked trend (toward, say, zero or one). Of course, the swings in the series will be pointed out, including rather persistent ones. The main facts we go over in this section, moreover, emphasize the longer run; short-run facts are discussed in more detail later. The growth facts we will focus most on are from the United States, but we will

show some data from other countries as well. They are, for the most part, referred to as the Kaldor facts, but there is no strict adherence here to the facts originally pointed to in Kaldor (1957).

2.1.1 Output grows steadily

One of the most remarkable facts in economics is the steady growth of output over the last centuries. The path for (the logarithm of) real U.S. output is shown in Figure 2.1.[1] The figure reveals almost constant growth over more than a century and the swings up and down seem minor from a bird's-eye perspective.[2] The notable exception is the Great Depression episode and the rebound after that, but after that hick-up the economy lands on "the same" growth path again. Thus, our regular business-cycle movements, including the most notable recent recession (the Great Recession, 2007–2009), are barely visible.

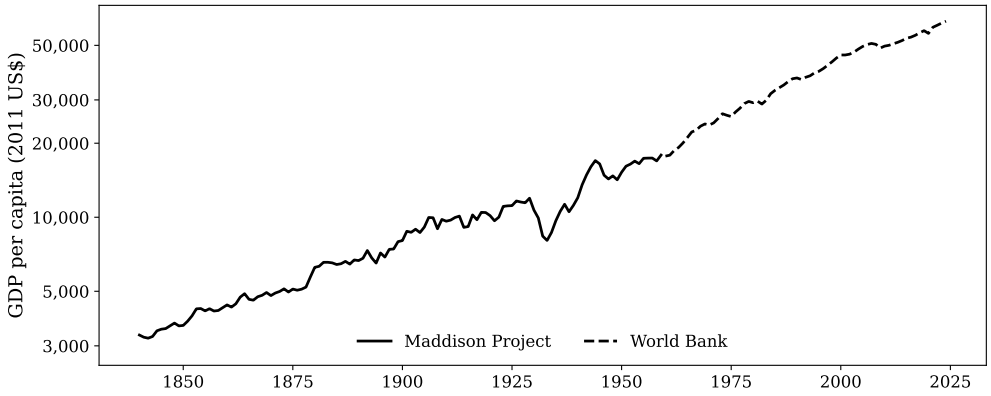


Figure 2.1: GDP per-capita in the U.S.

Notes: The figure plots GDP per-capita in 2011 prices in the U.S. 1840-2018.
Source: Maddison project.

One of our key goals now is to try to "account" for output growth, i.e., to provide a theory that offers a deeper understanding of the remarkable fact in Figure 2.1. We do so by focusing on the production side, i.e., how the basic inputs into production have evolved over time. We also look at their prices.

2.1.2 The basic resources behind output—and their prices

We begin by looking at capital.

[1] As discussed in the measurement section, systematic measurement did not begin until about a third into the twentieth century. The Maddison database goes much further back and then output estimates are based on available time series for, e.g., production, employment, and prices.

[2] A linear fit suggests that the series is well approximated by a annual growth rate of 1.86 percent. (Such a linear fit obtains an R^2 of more then 0.98.)

Capital input

As we shall argue, the process of growth is driven, at least in part, by capital accumulation. Figure 2.2 shows the capital-output ratio in the U.S. since the late 1920s. Apart from a marked jaggedness early on—during the Great Depression especially—we also see clear stability at a value of around 3. The measure for capital here is the standard one: "accumulated investments, minus depreciation."

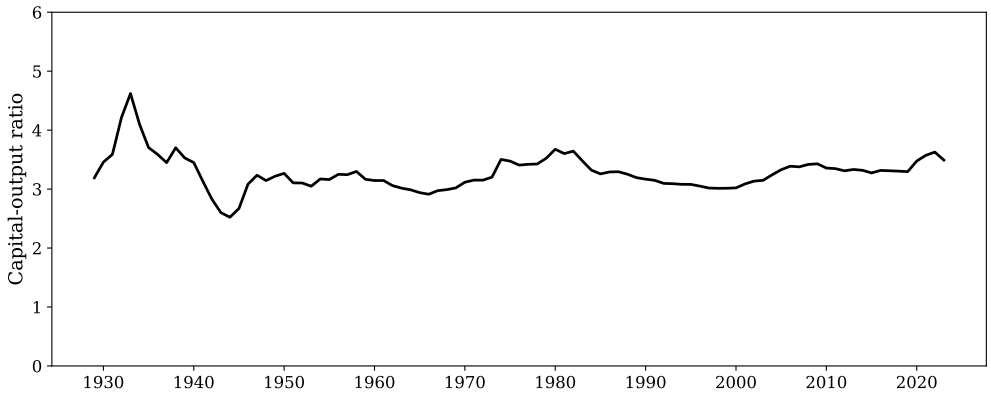


Figure 2.2: Capital-output ratio in the U.S., 1929-2022.

Source: FRED. Numerator: Current-Cost Net Stock of Fixed Assets and Consumer Durable Goods (K1WTOTL1ES000), Annual, Not Seasonally Adjusted, converted to billions of dollars. Denominator: Nominal GDP (GDPA), Annual, Not Seasonally Adjusted, reported in billions of dollars. The figure plots the ratio between fixed capital and consumer durables relative to the GDP.

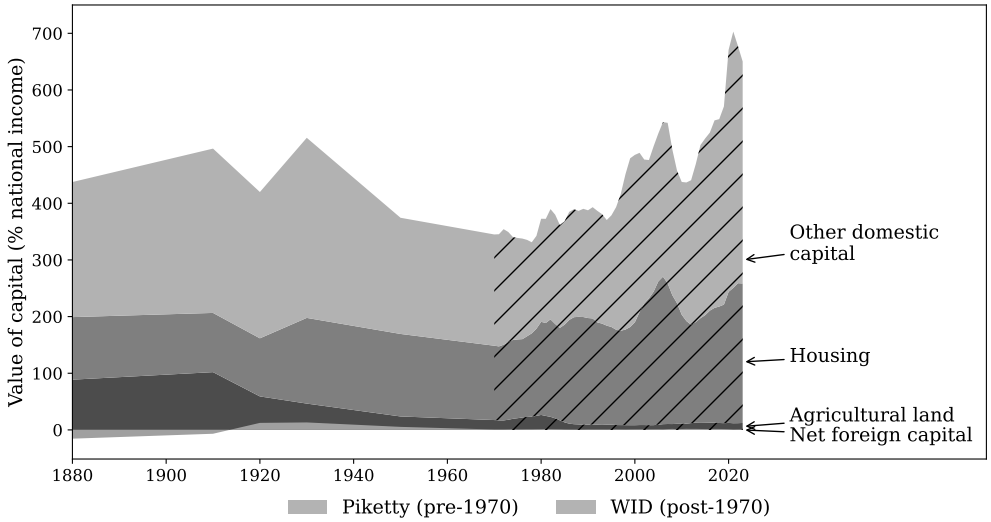


Figure 2.3: Wealth-output ratio in the U.S.

Source: Piketty (2014), Chapter 4, Figure 4.10[3].

[3]Data can be downloaded from http://piketty.pse.ens.fr/files/capital21c/en/xls/.

The focus here is capital as an input into production. It is nevertheless interesting to note that a broader interpretation of capital is wealth, which would include the value of land, housing, and so on. In Figure 2.3, we show the wealth-output data as computed in Piketty (2014).[4] We see marked stability again, though large changes in the composition of the capital stock, toward manufacturing capital and, especially, housing, and a total that is 4–5 rather than 3.

The price paid for using capital

How expensive has it been for producers to use capital over time? Most commonly, firms buy capital and use it until they scrap it, or sell it in market for used capital goods. It has become increasingly common for firms to instead rent capital (such as machinery or buildings), in which case the price paid for the use of capital is clearly the rent. But due to lacking systematic historical data on rents, measures of the cost of capital are instead constructed based on (a minimum of) theory. One way is thus to look at a measure of the returns on investments: on the margin, under competitive markets, this return should equal the marginal cost of the investment: the price we are looking for. Thus, we can look at stock-market returns as one measure of capital's price. Using binned data on stock-market returns—a return to investments in the capital of firms whose shares are publicly traded—from three long time-periods, Figure 2.4 shows no strong trends. If one were to look at shorter time periods, the variations in the stock market returns are of course very noticeable and large. These fluctuations are likely due to changes in asset valuations more than to changes in costs, which is why longer-run averages seem more appropriate.

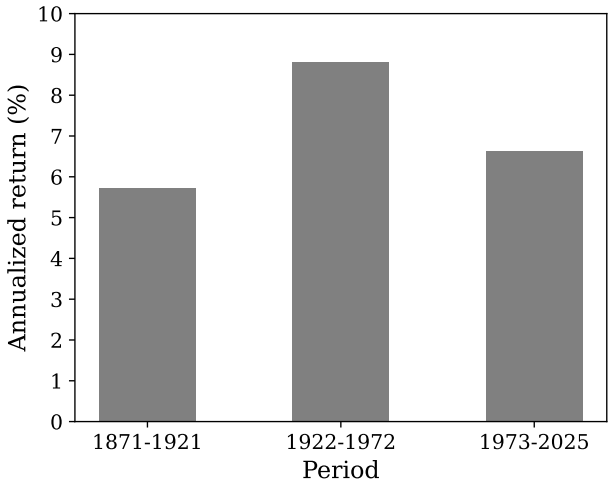


Figure 2.4: Return on capital.

Source: Shiller (2000). Data is annual, geometrically compounded returns to U.S. stock markets[5].

Alternatively, one can attempt to measure the cost side, but also using theory. The

[4] We use his data starting 1880.

[5] Data can be downloaded from http://www.econ.yale.edu/~shiller/data.htm.

"user cost of capital" is based on the foregone return to saving: by buying and owning capital, a firm is losing the return it would have received by merely saving the money. This measure also takes into account depreciation: a part of the capital is lost by using it, or needs maintenance to be kept in good shape. Moreover, it takes into account capital's change in value over time; computers, while not physically depreciating, lose value on the used market since new computers always out-compete old computers. Thus, the user cost is, roughly speaking, a market interest rate plus depreciation plus a the fall in value.[6] The user cost is also stationary, but of course also includes short-run swings

Labor input

The second major input into production is labor. Employment is one measure of input, but it is often relevant to take into account how many hours each employee works given that hours worked vary widely and many people have more than one job. Hence, a common measure of labor input is hours worked (in the marketplace) per adult. Various measures are available and we will show two here. First, Figure 2.5 shows that, since the beginning of the last century, hours worked per week have fallen, from around 28 hours to around 23 hours. Looking more closely at the graph, we see that since the end of World War II, hours look rather stable, without a net downward trend. This is a fact that is often referred to—that U.S. hours are stationary—but actually only accurate over the postwar period. Second, we see very large departures from trend in the figure; during the Great Depression, extreme unemployment rates account for the low hours, with a subsequent war-related upswing. In the U.S., unemployment movements, which are large, account for a big share of the movements in hours.

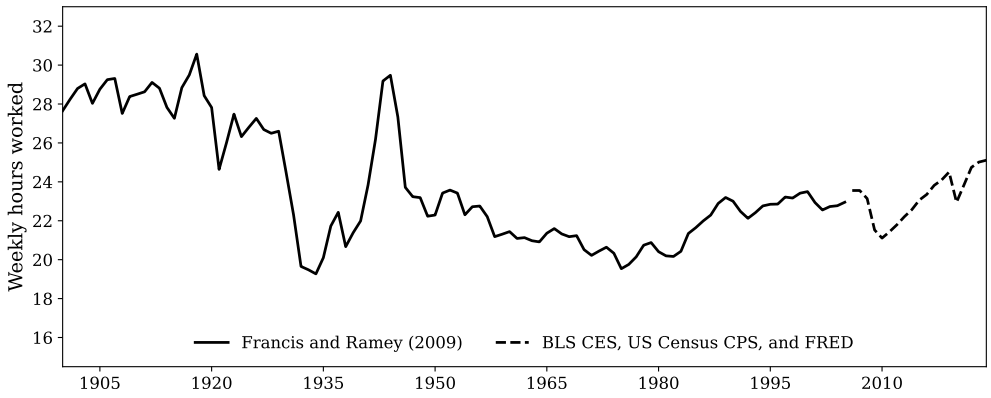


Figure 2.5: Average weekly hours worked in the U.S. (age 14+)

Source: Ramey and Francis (2009)[7].

Figure 2.6 shows hours worked over a longer time period, along with average real wages. Here, the downward trend in hours is even clearer (the graph depicts hours per employed, so the numbers are overall higher and do not account for changes in participation). The

[6]Other factors can appear in a user-cost formula, such as the role taxation plays for capital income an in deductions for depreciation.

[7]Data can be downloaded from https://econweb.ucsd.edu/~ vramey/research/Century_Public_Data.xls.

cumulative decline is almost 50%.

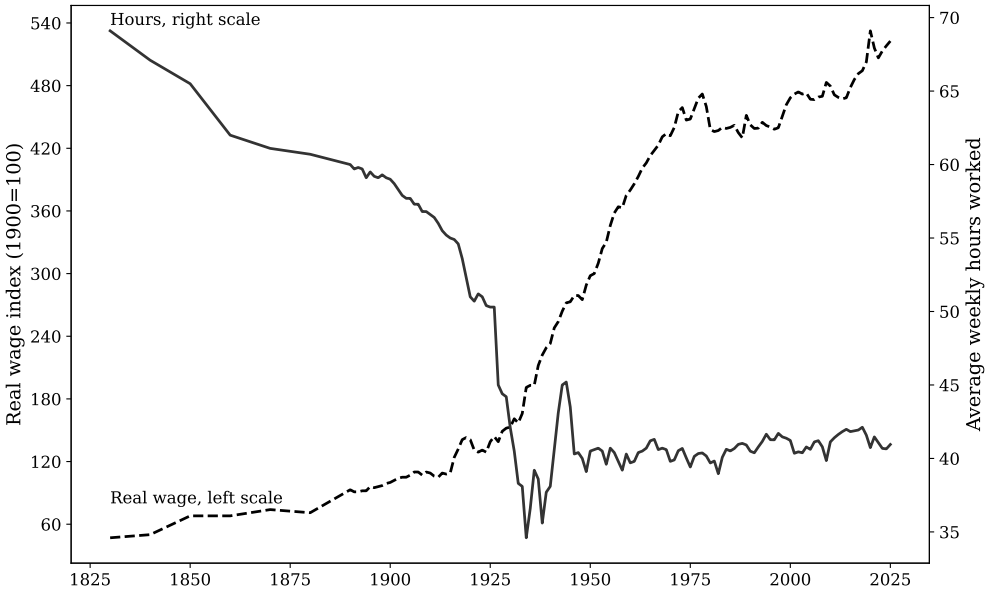


Figure 2.6: Average weekly hours worked in the U.S. (manufacturing)

Sources: Period 1830-1880: Whaples (1990), Table 2.1. Period 1890-1970: Bureau (1975) Series 765 and 803. Period 1970-2023: FRED, monthly series AWHMAN, annualized. Wage data source: Period 1830-1888: Williamson (1995) Table A1.1, (1900=100). Period 1890-2023: FRED, quarterly series LES1252881600Q, annualized (1982-84 CPI Adjusted Dollars) **Note**: Converted wage index in 1982-84 dollars from FRED to 1900=100 index by multiplying the FRED series by the ratio of wage index in 1983 from Williamson (1995) with wage index in 1983 from the FRED.

Wages

Figure 2.6 also shows real wages. Wages have risen at a remarkable average rate of around (or even above) the rate of output, with an exception in the most recent period. The secular increase is not surprising in some general sense: the standards of living have, slowly but surely, risen steadily for most Americans, and the major source of this rise is coming from higher and higher earnings. Given that hours are not showing an upward trend—in fact, the opposite—it must be that real wages have risen steadily. There are also some movements in how total output is divided up into capital and labor income; we will discuss these later, but the first-order aspect here is that the shares have been quite stable.

2.1.3 Taking stock: a "neoclassical" picture emerges

The data on capital and output had puzzled economists; in particular, the constancy of the capital-output ratio at a value around 3 suggested something quite stark—it suggested a rigid technology structure where labor played no role and capital and output were always in the same proportions—and this was hard to square with how we knew that production took place in practice.

Solow (1956) however, took a less immediate production perspective on these facts. He noted what you saw in the previous two sections: there has been a steady rise in output, an equally steady rise in capital with stationary, or even declining, hours worked. At the same time, the price of capital has been stationary while wages have had a significant upward trend. These facts did not, by themselves, appear mysterious. Solow in particular noted that it is natural that as the price of an input rises, the use of the input declines. More precisely, from the perspective of production theory, at a higher relative input price, a firm uses less of that input relative to other inputs. Labor has become more and more expensive relative to capital, and its use has fallen, again relative to capital: the capital-labor ratio has risen at the rate of output growth, or even slightly more.

The other side of the coin, which was important to address, is about accounting for these changes in relative input prices: what made labor more and more expensive relative to capital? One natural explanation would have its roots in technological change. In particular, suppose it is directed toward labor: labor becomes more productive over time per unit of hour worked. If capital and labor are complementary, this factor would lead to an increased value of capital on the margin. However, as we saw, the market return to capital has remained stationary—it has not risen. Solow noted, however, that the stationarity could be explained by *neoclassical* forces. A production function that has decreasing marginal returns to each input is labeled neoclassical and with such a production function, as the relative amount of capital—capital used per worker—rises, the marginal value of capital would fall. If firms buy inputs in competitive input markets, moreover, this would be reflected directly in the market return to capital. Thus, a story emerges where technological change directed toward labor, along with neoclassical forces, can, at least potentially, account for the historical data on relative input quantities and relative input prices.

The neoclassical features led Solow to describe *aggregate* output as being generated by an *aggregate* production function, with *aggregate* capital and *aggregate* labor as inputs. Solow also posited, along the lines above, that the production function likely changed nature over time—that there was some technological progress—and built a model around these ideas; we will briefly describe this model momentarily but let us first focus on how, given Solow's perspective, one could take the next natural step: accounting for the sources of aggregate growth as coming from growth in inputs and growth in technology. This accounting procedure, along with the development of Solow's neoclassical growth model, would allow us to obtain a much less mysterious, and in fact quite natural and operational, account of how output grows over time. As a matter of microeconomic theory, the existence of an aggregate production function, i.e., a functional mapping from the total (economy-wide) quantities of inputs only into some measure of aggregate output, is not easy to establish in general and has also, as the box below discusses, been subject to some controversy.

The existence of an aggregate production function

There are assumptions under which a functional relationship can be established between input quantities and a price-independent output measure. However, these assumptions are extremely specific, indeed knife-edge, cases. Think of a static economy producing two goods, x_1 and x_2, both from capital and labor but with different production functions;

also imagine that there are no restrictions on how inputs can be allocated across the two sectors. Then if a relative price between x_1 and x_2 is assumed—let us call it p—it is straightforward to see that a competitive equilibrium will generate a mapping between the total input quantities and the value of total output: perfectly competitive markets would allocate, or a planner could equivalently allocate, capital and labor so as to maximize output. As the total amounts of capital and labor would vary, total output would change. Thus, we would obtain a mapping from inputs to output. But this mapping would nontrivially involve p: it would not be a pure production-function relationship. Thus, one would need to add a demand side, thus endogenizing p, to obtain a pure mapping from input quantities to a measure of output. But then preferences (or whatever gives rise to demand) need to be described, and they will generally, as would p in a direct sense, influence the mapping. Thus, a pure production mapping is hard to imagine.[a]

The existence and usefulness of an aggregate production function was hotly debated in the so-called Cambridge capital controversy during the 1950s. This controversy, which had its two head quarters in the two well-known Cambridges (the University of Cambridge, U.K., chiefly represented by Joan Robinson and Piero Sraffa, vs. Paul Samuelson and Robert Solow at MIT, Cambridge, Mass., U.S.), also involved the notion of "aggregate capital," but in essence it focused on the aggregate production function. The controversy on the existence conditions can be summarized as having been won by Cambridge, England, whereas the usefulness controversy was arguably won by Cambridge, Mass. With the tools of modern macroeconomics, one can solve large models and see to what extent departures from aggregation play an important role quantitatively. Some such endeavors have been undertaken and point to limited departures, but no fully systematic analyses have been conducted yet.

[a] If you assume that the two production functions have identical isoquants—one is a scalar multiplication of the other—then it is possible to construct the mapping, as a relative price p is implied from the production technologies alone. Try to show this as an exercise!

2.1.4 Growth accounting

Solow (1956) introduced "growth accounting" as a way to implement the notion that one could break down aggregate output growth into sub-components. With the use of a rather limited amount of theory, Solow was thus able to quantify the relative importance of different sources of U.S. growth.

Solow's growth accounting made the following assumptions: (i) aggregate output Y_t is generated from an aggregate production function $F_t(K_t, L_t)$, where K_t is aggregate capital input, L_t aggregate hours worked, and the subscript t denotes that technological change may move the production function upwards over time; (ii) F has constant returns to scale (CRS) and neoclassical properties; (iii) there is perfect competition for inputs and, hence, firms maximize profits. The CRS assumption, we know from microeconomics, leads to zero pure profits in a perfectly-competitive equilibrium, which Solow thought was a good approximation and, besides, being able to replicate production would at least ensure that if all inputs double, output would double, so less than constant returns to scale was not thought of as reasonable.

Throughout the text (when not otherwise noted), we will use the convention that lower-case letters are in per-capita real terms. Thus, in this chapter we use y_t to denote per-capita output, i.e., Y_t divided by the size of the population; denoting population by N_t, we have $y_t = Y_t/N_t$. Note that, since F_t is CRS, we can write $y_t = F_t(k_t, \ell_t)$, where k and ℓ are thus capital and hours worked per-capita terms. Most of the time in this chapter, beginning here and now, we will also abstract from population growth and simply consider a population of constant size ($N_t = N$). It is sometimes convenient to normalize the population size to 1 so that we have $y_t = Y_t$.

We will now derive Solow's growth-accounting equation and for this we will use a continuous-time formulation where all variables are functions of time: we will write $y(t) = F(k(t), \ell(t), t)$. We will assume differentiability of these functions of time, so that

$$dy = \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial k}dk + \frac{\partial F}{\partial \ell}d\ell = \frac{\partial F}{\partial t}dt + \frac{\partial F}{\partial k}k\frac{dk}{k} + \frac{\partial F}{\partial \ell}\ell\frac{d\ell}{\ell},$$

where all the partials are evaluated at $(t, k(t), \ell(t)$. Dividing by output, using r and w to denote capital's rental rate and the wage (the prices, as well as the quantities, also depend on time but this dependence is omitted for notational convenience), and then using firm profit maximization, we obtain

$$\frac{dy}{y} = \frac{\partial F}{\partial t}\frac{dt}{y} + \frac{rk}{y}\frac{dk}{k} + \frac{w\ell}{y}\frac{d\ell}{\ell}.$$

Here, r and w (both with time dependence suppressed) replaced the two marginal products: this is what follows from taking first-order conditions of the profit maximization problem

$$\max_{k,\ell} F(k, \ell, t) - r(t)k - w(t)\ell.$$

We finally let $1 - \alpha$ denote labor's share of income, which we know from NIPA; α here of course may also depend on time. Then the CRS assumption gives us the capital share as α: total input costs equal output, thus delivering zero profits. This equation allows us to obtain an account of the sources of growth:

$$\frac{dy}{y} = \frac{\partial F}{\partial t}\frac{dt}{F} + \alpha\frac{dk}{k} + (1 - \alpha)\frac{d\ell}{\ell}. \tag{2.1}$$

The term $(\partial F/\partial t)(dt/F)$ is labeled the *Solow residual*, because it can be calculated as a residual of the growth in output that cannot be accounted for by the growth in capital and labor (the second and the third term of the right-hand side). We see that output growth is a weighted sum (over small intervals in time, as we have used derivatives) of capital input growth and labor input growth, where the weights are their respective income, or cost, shares, plus the Solow residual, which is a measure of how the production possibility frontier has moved out over time.

The Solow residual in (2.1) expresses the direct effect of time on production (in percentage terms). In general, this effect depends on the input pair (k, ℓ) at which F is evaluated. If one makes further assumptions on how technology shifts the production function, it is however possible to derive specific series for technological change that are independent of

the economy's current capital-labor mix. We will discuss two such assumptions because they are commonly used; they are by no means the only ones imaginable, but especially the second one will play a key role later.

One assumption is to let $F(k, \ell, t)$ take the form $zF(k, \ell)$, where F is a time-independent function and only z depends on time. In this case, equation (2.1) can be expressed as

$$\frac{dy}{y} = \frac{dz}{z} + \alpha\frac{dk}{k} + (1-\alpha)\frac{d\ell}{\ell}. \tag{2.2}$$

Here, z is TFP: total factor productivity. It can equivalently be thought of as a common, "Hicks-neutral" factor multiplying both inputs.[8]

Alternatively, we can make the assumption that technology is *labor-augmenting*: we define $F(k, \ell, t) = F(k, z\ell)$, again with a time-independent function F. Now z has a different meaning. In this case, since $(\partial F/\partial t)dt = (\partial F/\partial \ell)\ell dz$, we can write the original growth-accounting equation (2.1) as

$$\frac{dy}{y} = (1-\alpha)\frac{dz}{z} + \alpha\frac{dk}{k} + (1-\alpha)\frac{d\ell}{\ell}, \tag{2.3}$$

after having replaced $\partial F/\partial \ell$ by w/z and recognized that $w\ell = (1-\alpha)F$.[9]

We now show some results from growth accounting, implemented the way Solow came up with it. In practice, it is important to take into account how the quality of the inputs change over time; in particular, one may want to adjust labor input to account for human capital accumulation, or else part of the Solow residual will reflect the increasing quality of this input. Also, since growth accounting in practice is carried out using data from discrete time periods, like years, one must be clear on which input shares to use, e.g., taking an average over the shares in t and $t+1$ when accounting for growth between these two years.

We first look at the traditional measure of productivity: labor productivity. Figure 2.7 shows an average growth rate of a little below two percent per year, with significant ups and downs; currently, we are in a down period.

Figure 2.8 shows a full time series (along with those for some other countries) over a longer time period. These series are smoothed.

We thus see stable, positive labor productivity growth, hovering around two and a half percent per year. How about the growth in total factor productivity? Figure 2.9 gives the answer from two sources, now in un-smoothed form. Here too, we see growth at a little below two percent per year, with significant movements up and down that we will return to later.

Figure 2.10 shows smoothed data for TFP growth over a longer time period. The patterns are similar.

2.1.5 The dynamic system

We are now equipped with measures of output and input aggregates, as well as with a measure of aggregate technology, in the form of TFP. Solow's next step was to use his framework to

[8] This is true since we have assumed that F is homogeneous of degree 1 in (k, ℓ).

[9] Similarly, one could define a capital-augmenting technology series by letting the z multiply capital.

[10] Gordon defines "labor productivity as real GDP divided by an unpublished quarterly BLS series on hours for the total economy, including the private economy, government, and institutions."

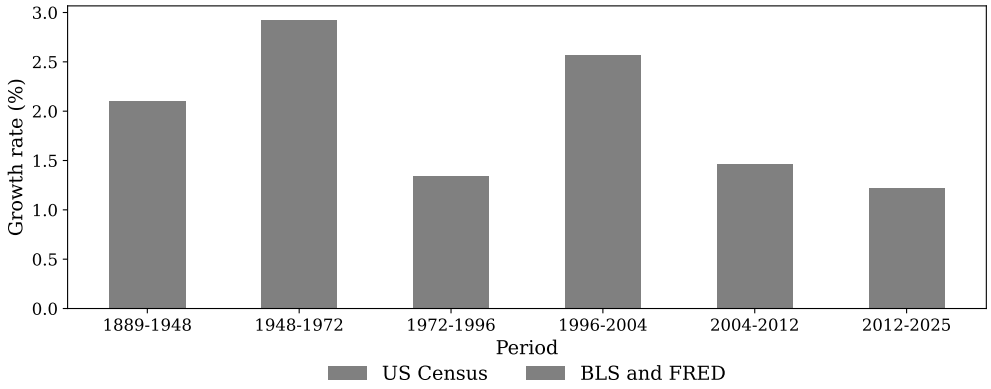


Figure 2.7: Labor productivity in the U.S., sub-periods

Source: Period 1889-1948: Bureau (1975), Part II, Series W 1, pp. 948.. Period 1948-2023: U.S. Bureau of Labor Statistics and FRED Quarterly total economy hours worked (in billions of hours) series are from BLS "Hours Worked in Total U.S. Economy and Subsectors." Quarterly real GDP (in billions of 2017 dollars) is from FRED Real gross domestic product (GDPC1).
Note 1: Data constructed following Gordon (2012)[10] Percentage logarithmic growth rates are calculated between the first quarter of each of the listed years, e.g., 1948:Q1 to 1972:Q1. To extend the series back from 1948 to 1891, annual NIPA data on real GDP prior to 1929 are ratio-linked to the real GDP data of Balke and Gordon (1989), and the BLS hours data prior to 1948 are ratio-linked to the man-hours data of Kendrick (1961) (see pp. 330-32)." Unlike Gordon (2012), to extend the series back from 1948, we used Census Bureau Hisitorical Statistics. Then the series for both periods are then re-indexed to 1948=100.

probe further into the mystery of the all but constant capital-output ratio.

Step one in this endeavor was to explicitly link time periods (years) by noting that tomorrow's capital aggregate stock is today's stock, plus new investment minus the part of capital that has depreciated.[11] With a constant rate of capital depreciation δ, this yields

$$k_{t+1} = (1-\delta)k_t + i_t. \tag{2.4}$$

We see an equation that is consistent with the long-run data, if all the variables appearing in it grow at a common rate. We can rewrite the above equation as

$$\frac{k_{t+1}}{y_{t+1}}\frac{y_{t+1}}{y_t} = (1-\delta)\frac{k_t}{y_t} + \frac{i_t}{y_t}$$

and denoting the (net) growth rate of output as γ_t, this equation can be expressed as

$$\frac{k_{t+1}}{y_{t+1}}(1+\gamma_t) = (1-\delta)\frac{k_t}{y_t} + \frac{i_t}{y_t}.$$

Clearly, if capital, investment, and output all grow at a same constant rate γ and i_t/y_t is equal to a constant value s, then this equation is consistent with capital-output ratio that

[11] A similar accumulation equation can be formulated for human capital. We delay this discussion until our Chapter 13 on growth.

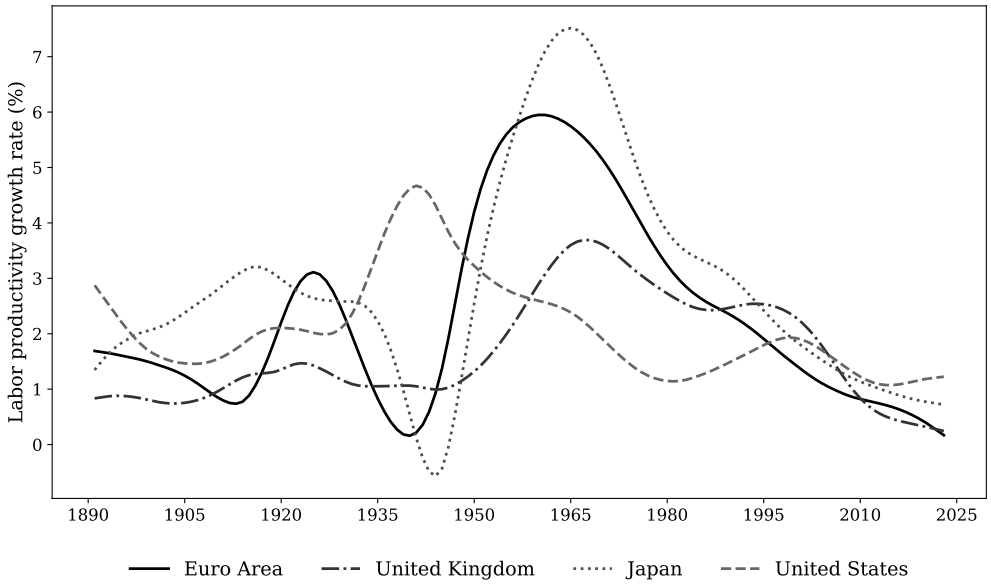


Figure 2.8: Labor productivity for a selection of countries

Note: Hodrick-Prescott-filtered annual growth of labor productivity per hours worked. Following Bergeaud et al. (2016), we focus on 30-year cycles, which implies an HP-filter value of 500 for lambda.

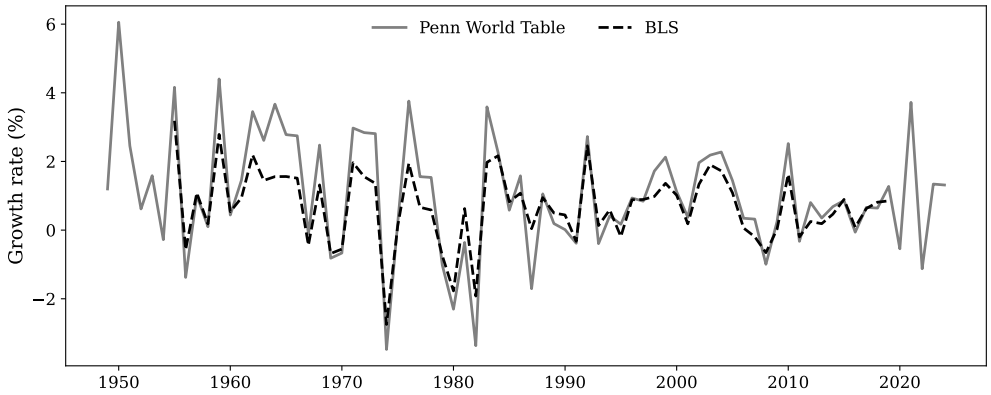


Figure 2.9: TFP in the U.S., two measures

Sources: Series 1: FRED, "Total Factor Productivity at Constant National Prices for United States (RTFPNAUSA632NRUG)," reported by Penn World Tables. Available for 1955-2019 Series 2: Utilization-adjusted quarterly TFP series for the U.S. Business Sector, 1948-2022 from Fernald (2012).

does not change over time:

$$\frac{k_t}{y_t} = \frac{s}{\gamma + \delta}$$

for all t. The investment-output ratio is not constant over time; in particular, it fluctuates significantly. The consumption-output ratio for the U.S. is plotted in Figure 2.11; here, consumption is defined as private plus government. One minus this measure is close to the

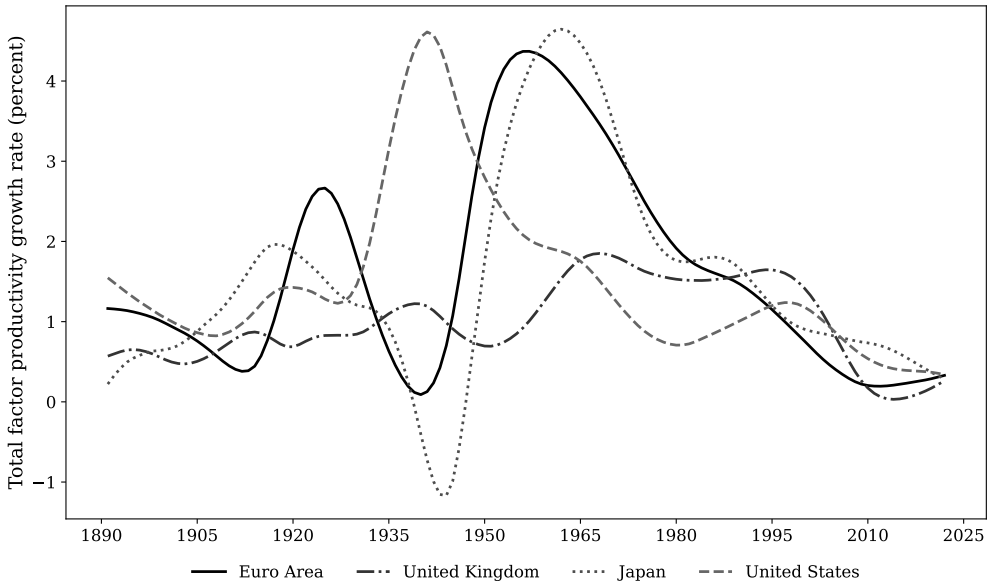


Figure 2.10: Historical TFP for a broader set of countries

Source: Following Bergeaud et al. (2016), we focus on 30-year cycles, which implies an HP-filter value of 500 for lambda. TFP growth rate for years after 2012 is from the OECD data series Multifactor productivity. Aggregate TFP growth rate for EU19 countries is calculated by taking a weighted average of the growth rates of each country where weights are the share of each country in the total GDP of the EU19 in each year. Data obtained from OECD Gross domestic product (GDP).

ratio of investment (again, private plus government) to output, since net exports are near zero in the U.S. as a fraction of GDP.[12] We see significant movements in s early in the 20th century and thereafter small movements, possibly with a slight downward trend. But the overall assumption of a constant investment, or saving, rate appears to be a good one.

The above discussion allows a mechanical account of how one can interpret the capital-labor ratio: it is a simple function of the rate of saving, the economy's growth rate, and the rate of depreciation of capital, along what has been labeled the *balanced growth path*. The numerical values can be squared, too: with a depreciation rate of around 0.08 and a net growth rate of around 0.02, a saving rate of 0.30 delivers a capital-output ratio of 3.[13]

However, this account still does not give an answer to the question why: why is the economy always (almost) at this value? That is, it explains if the capital-output ratio is 3 at some point in time, it will remain 3. But why is it 3 to start with? Solow found an answer.

Solow considered the following dynamic system, which is the logical implication of the above reasoning:

$$k_{t+1} = sF(k_t, (1+\gamma)^t \ell) + (1-\delta)k_t$$

for all t. This system is almost exactly what we have looked at before. First, we have replaced

[12]This comes from the national accounting identity $Y = C + I + NX$, where C and I both contain private as well as government expenditures.

[13]The saving rate of 30% is larger than what is implied in Figure 2.11. This discrepancy is largely because Figure 2.11 includes the government spending in consumption.

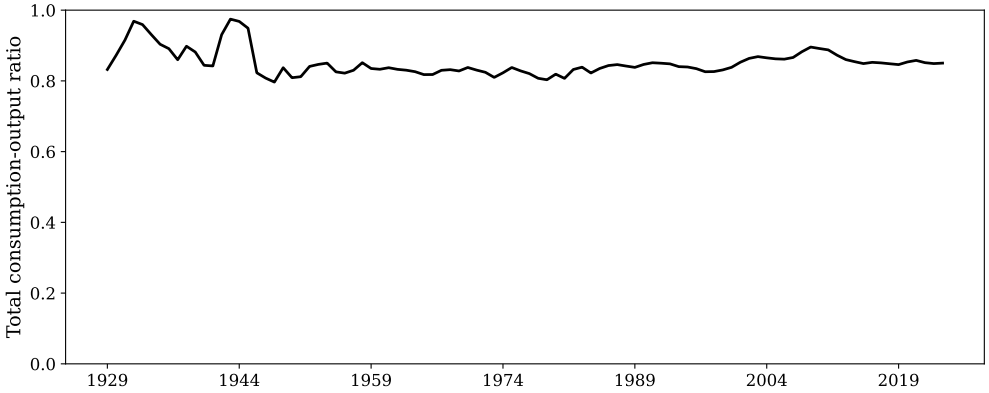


Figure 2.11: Ratio of total consumption to output

Source: BEA, NIPA Table 1.1.5 Calculated as the ratio of two series: Numerator: Sum of Personal consumption expenditures (DPCERC) and Government consumption expenditures and gross investment (A822RC), Annual, Millions of dollars Denominator: Gross domestic product (A191RC), Annual, Millions of dollars The figure plots the ratio between total consumption expenditures (sum of all goods and services) relative to the GDP.

i by sy, since we take saving to be a constant fraction of output, in consistency with the above evidence. Second, there are two additional assumptions: we have set hours worked to a constant, ℓ, and we have made technology growth appear only in a *labor-augmenting* form, i.e., technical change is equivalent to raising labor input, or the quality of labor input. This was based on the hunch above: if labor units become more and more productive due to technological change, the return to capital does not have to fall due to a higher ratio of capital to hours worked. It turns out that this assumption is actually crucial. Namely, Uzawa (1961) proved that, for a production function of two inputs where one is limited—labor hours, in this case, are constant—to admit exact balanced growth in a model of this kind, technological change has to take this form.[14] So we can obtain balanced growth if and only if the above assumptions are met; whether labor input is constant or declining does not matter.

The dynamic system can be rewritten

$$(1+\gamma)\tilde{k}_{t+1} = sF(\tilde{k}_t, \ell) + (1-\delta)\tilde{k}_t;$$

the equation is obtained by means of a simple variable transformation—$\tilde{k}_t \equiv k_t/(1+\gamma)^t$, a stationary variable if k grows at the net rate γ—and division by $(1+\gamma)^t$.

Now note, first, that there is a constant solution to this system: $\bar{\tilde{k}}$. It is the unique value that (under some minimal conditions) solves

$$(1+\gamma)\bar{\tilde{k}} = sF(\bar{\tilde{k}}, \ell) + (1-\delta)\bar{\tilde{k}}.$$

So in a situation where $\tilde{k}_0 = \bar{\tilde{k}}$, we obtain $\tilde{k}_1 = \tilde{k}_2 = \cdots = \bar{\tilde{k}}$. That is, capital grows at a

[14]The proof of Uzawa's important theorem can be found in Appendix 3.A, which discusses the Solow model in detail.

constant rate γ, since $k_t = \tilde{k}_t(1+\gamma)^t$. The same is then true for investment and output, since $y_t = F(k_t, (1+\gamma)^t\ell) = (1+\gamma)^t F(\tilde{\bar{k}}, \ell)$ follows when F is CRS.

We have established that if the initial capital stock has a particular value, this economy will grow at a constant rate. However, the second, and most remarkable, thing to note about this system is a convergence property. Solow showed, under very weak conditions, that for any given initial condition on capital, the ensuing capital sequence, and hence the economy's aggregate variables will *converge* to the balanced growth path. We will explain this in detail in the next chapter. Intuitively, it is the neoclassical production function—the very feature that was used to make sense of why a higher input price is consistent with lower input use, and at the same time, under competition, a higher marginal productivity—that explains convergence: when capital is comparatively low, growth is comparatively fast, because its marginal productivity is high, delivering higher capital accumulation per unit of capital. Conversely, when capital is high, its growth rate is low, so there is movement back toward the balanced growth rate no matter where the initial capital stock is.

The neoclassical convergence mechanism de-mystifies why the value of the total capital stock on average is worth roughly 3 times annual output: it doesn't have to be exactly 3 at all times, but deviations bring it back toward 3. It also produces a dynamic framework that, as we shall see in the rest of the textbook, has become the workhorse model for macroeconomics, much because it offers a coherent account of how the macroeconomic aggregates have evolved historically. The analysis of business cycles, for example, builds on the neoclassical framework with various stochastic shocks added to the dynamic system. These shocks could be "supply shocks" or "demand shocks," but they have in common that their *propagation* through the economy—how the macroeconomic variables respond in the short run, which can differ greatly across different kinds of shocks—eventually takes us back toward the balanced growth path. This is ensured by the convergence property of the system.

The final, unresolved, issue is that some of our assumptions above are mere mechanical descriptions of the data: investment is a constant rate s of output, and hours worked are constant at ℓ. In the real economy, these two features should be the results of conscious choices made by households. Consumption choices are made continuously, and people can influence how much they work. Moreover, a theory that adds consumer choice will also allow us to make welfare statements, which the analysis so far does not. We will turn to households' choices momentarily: we will "rationalize" the s and the ℓ based on microeconomic theory—utility maximization. However, let us first briefly revisit the movements of input shares.

2.1.6 Input shares

On a balanced growth path, the shares of income paid to capital and to workers, respectively, are constant. This is because in the theory Solow proposed, rk grows at the rate of output (r is constant but k grows at the rate of output) and $w\ell$ does as well (w grows at the rate of output and ℓ is constant, at least over the postwar period). Thus, as shares of output, they should be constant. Have they been? We see in Figure 2.12 that there is rather remarkable constancy over time.

In the most recent years, a downward trend of the labor share can been noticed, however, as is evident if the time interval is restricted. Figure 2.13 illustrates, for a few developed countries, that the labor share has been declining. Figure 2.14 shows the same fact as a global

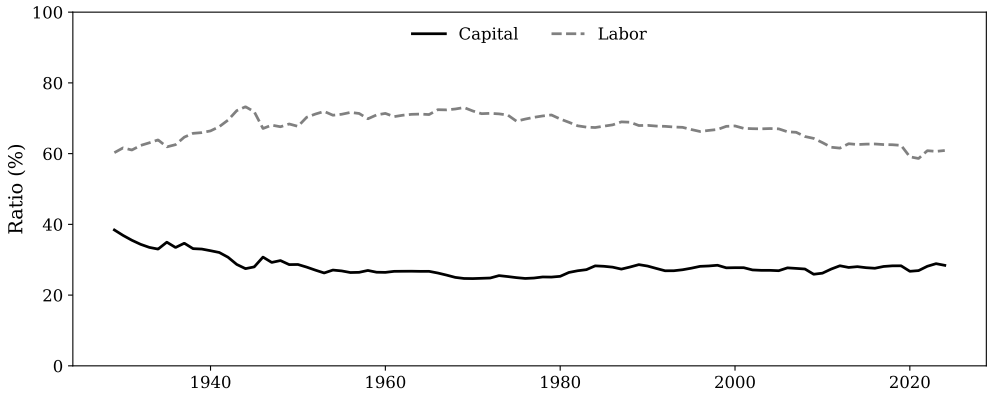


Figure 2.12: U.S. factor shares over time

Source: NIPA Table 2.1 Factor shares are calculated as compensation of employees and capital income divided by personal income. Capital income is calculated as proprietors income with inventory valuation and capital consumption adjustments plus rental income of persons with capital consumption adjustment plus personal income receipts on assets.

average. The downward trend has been subject to much scrutiny and research recently, but for now we will maintain the stylized fact as "the labor share is close to constant over time."

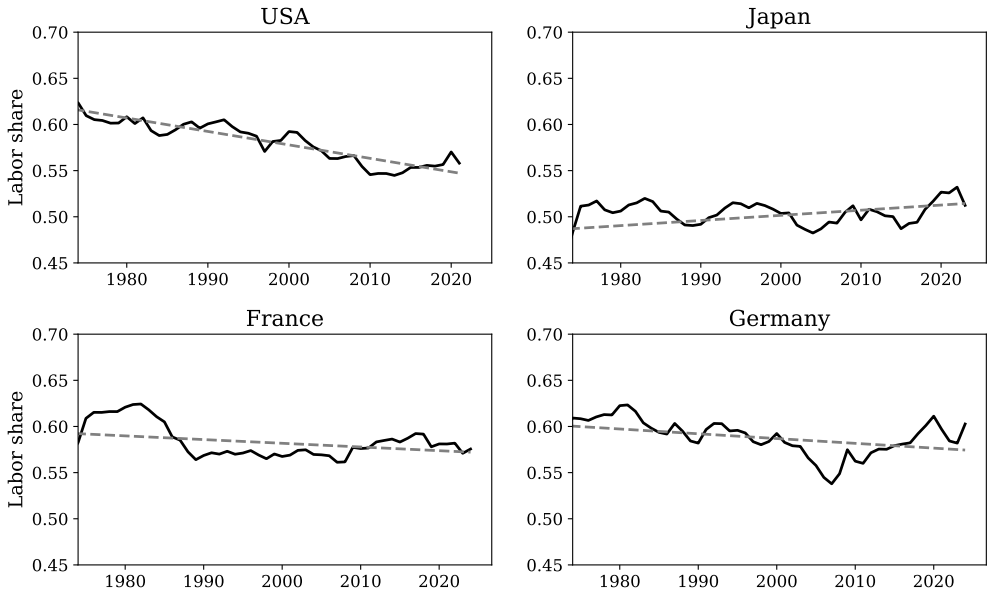


Figure 2.13: Labor share

Source: Karabarbounis and Neiman (2014). **Note**: No corporate labor share data is available for Japan, thus total labor share is plotted instead.

Virtually all applied macroeconomic studies employ an aggregate production function that is of the Cobb-Douglas variety, i.e., we have $F(k, \ell) = Ak^{\alpha}\ell^{1-\alpha}$, where α is constant over time. This function has the property that $F_k(k, \ell)k/F(k, \ell) = \alpha$, i.e., the income shares under perfect competition are independent of the values of k and ℓ. We see in the data that

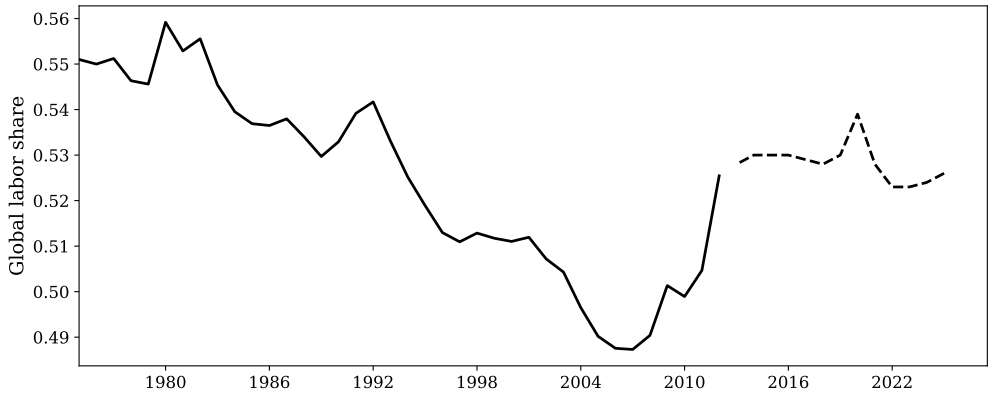


Figure 2.14: Global labor share

Source: Karabarbounis and Neiman (2014). **Note**: Global labor share of the corporate sector is the average of country-level labor shares weighted by corporate gross value added measured in U.S. dollars at market exchange rates. Global labor share of the overall economy is the average of country-level labor shares weighted by GDP measured in U.S. dollars at market exchange rates.

although the shares are not literally constant, their movements are relatively minor, even as economies go through recessions and booms. The Cobb-Douglas function thus conveniently generates a decent approximation to the data, which is why it is so often used.

2.1.7 Summing up

We summarize the main stylized (notice the repeated appearance of the word "roughly" in the descriptions below) facts:

1. output per capita has grown at a roughly constant rate
2. the capital-output ratio (where capital is measured using the perpetual inventory method based on past consumption foregone) has remained roughly constant
3. consumption as a fraction of output has been roughly constant
4. the wage rate has grown at a roughly constant rate equal to the growth rate of output
5. the real interest rate has been roughly constant, seen over a longer period of time
6. labor income as a share of output has remained roughly constant
7. hours worked per capita have been roughly constant over the recent half a century.

These facts are consistent with aggregates obeying a neoclassical structure whose core is a CRS production function with labor-augmenting technical change and decreasing marginal products in each input, constant labor supply, a constant rate of capital depreciation, and a constant investment-output (saving) ratio.

2.1.8 Rationalizing saving and labor-supply choices

We now discuss how macroeconomists theorize further to make sense of households' observed choices for saving and labor supply. The aim is to add a richer microeconomic structure that is qualitatively, and quantitatively, consistent with the data. The central object here will be utility functions and a key question is whether there are utility functions consistent with the observed choices. A follow-up question is whether, given such utility functions, the main dynamic properties—in particular convergence—will be maintained. We briefly address the first of these questions in the present chapter and postpone the second until a later chapter.

Before we introduce utility functions, we will briefly discuss the population structure, along with our general approach.

Time and people

We begin with time and then describe our people.

Time The dynamic system above was described in discrete time, i.e., time periods are integers. It is also possible to describe the system in continuous time (here, we would use t as an argument of our functions and not as a subscript, e.g., $y(t)$ vs. y_t). The main workhorse model developed in this book is using discrete time, mostly because we consider it somewhat easier to teach.[15] There is, however, no substantive difference between the two approaches and both are common in practice.

Second, we will almost always assume that time is infinite; the main exception is when we illustrate mechanisms in, say, two-period models. The reason for adopting infinite time is mostly practical but to some extent also conceptual. The same economic structures as we study under the infinite-horizon assumption can also be analyzed under the assumption that time is finite. However, then there is a built-in non-stationarity: time itself obtains importance—because it captures how far we are from "the end" and, certainly, if the end would be near, many decisions in the economy would change radically. So long as decisions and outcomes today are not more than marginally affected by the exact end date, it is simply more convenient to consider time to be infinite; then, time is not important in itself. Most macroeconomic models with an end date T have the feature that T is virtually irrelevant for decisions far from T (i.e., decisions at $t \ll T$); when the models we look at do not have this property, we will point it out.

People At any point in time, the economy is inhabited by individuals, or households, and our maintained assumption will be that they are utility maximizers. By assuming utility maximizers we mean we "rationalize" the choices we observe in the data by treating these choices as optimal given well-defined utility maximization problems. Thus, we back out how consumers must value consumption (and leisure) based on their own behavior. This approach is powerful, since it allows us to make statements that directly reflect the welfare

[15] The reason is that it is easier to make concrete; for example, dynamic optimization can be more straightforwardly connected to basic optimization theory in one or more (a finite number of) variables; when there are stochastic shocks, continuous time requires more investment still, and becomes a bit more abstract as well.

of the population. In the analysis of economic policy, being able to make welfare statements is crucial, since it allows us to compare policies on normative grounds.[16]

The most common macroeconomic model has a very stark, and highly stylized, description of the population: there are many identical individuals alive at any one point in time, and these individuals, moreover, live forever. "Identical" here means that they have the same utility functions and constraints and, therefore, face identical maximization problems. "Many" is important in that it allows us to make sense of price-taking behavior, but the number itself does not matter. "Live forever" sounds much sillier than it is: the notion here is that people are representatives in a dynasty, i.e., a consumer at time t values not only her own consumption but also that of her children, grandchildren, and so on. Given how people appear to care about their children and grand-children—given the resources they spend on them while alive and also in the form of bequests—it does seem a very natural starting point. It should be pointed out that the assumption here is not just that people care about their offspring: they are truly altruistic about their offspring. In other words, their preferences are aligned. We will touch on the possibility that they are not later on in the text.

Of course, macroeconomic models with more explicit and realistic structures are also common. First, building in a life-cycle pattern—since, at the very least, as a person ages, needs and abilities change—is common.[17] Then, the addition of children and altruism toward them would still deliver a model that in its macroeconomic features is quite similar to the simpler model. If life-cycle models do not have altruism toward children, such as in the simple "overlapping-generations model," they can (but do not necessarily) behave differently than than dynastic models. We will discuss this in some detail too.

Second, nowadays a "heterogeneous-agent" framework has been developed and become a very commonly used workhorse for macroeconomic analysis. In it, households are different in a variety of ways (income, wealth, preferences, etc.). Such models share many properties with the simplest dynastic setting, but of course also add richness. One aspect of these models is that they allow us to jointly study macroeconomics and inequality. Another one is that they behave quite differently (and, arguably, in ways closer aligned with the data) in some respects, in particular in terms of how the economy responds to various shocks and to policy changes. From our perspective here, however, the key observation is that heterogeneous-agent models can be viewed as extensions of the dynastic representative-agent setting rather than as fundamentally different. Third, some macroeconomic models also have richer models of the household structure, explicitly incorporating couples and children. This is, so far, a smaller literature, however.

Preferences

Households will be assumed to have utility functions that are time-additive, with consumption in periods t and $t+1$ evaluated as $u(c_t) + \beta u(c_{t+1})$, where u is strictly increasing and

[16] There is a strand of macroeconomic literature that builds in behavioral elements; after all, the literal interpretation of the data that consumers make perfect decisions at all time is of course very strong. One interpretation of this research is as a robustness check: if minor departures from rationality create major changes in the results, we should perhaps worry. This research is ongoing and we have no systematic treatment of it in this text.

[17] Death can then be modeled as deterministic or stochastic.

strictly concave. Thus, the same function u is used for consumption in both periods, but there is a weight β on $t+1$ consumption. The fact that u is the same for both consumption goods implies that consumption in both periods are normal goods: with more income, consumers would like to consume more of both goods, which seems very reasonable in this application.[18] Another aspect of this setting is an element of *consumption smoothing*: there is decreasing marginal utility to consumption in both periods so spending all of an income increase in one period will in general never be optimal. Furthermore, $\beta < 1$ captures impatience, or a probability of death—or any other reason for down-weighting future utility—and will be a typical assumption.

Choice

An important part of the text will explain intertemporal choice from first principles: different methods for solving intertemporal problems, with and without uncertainty, along with a number of important macroeconomic applications. Here, the purpose is to very briefly explain the key steps, heuristically, in order to account for the growth facts.

Conceptually, the way consumers make decisions—if able to choose when to consume their income—is according to basic microeconomic principles: so as to set their marginal rate of substitution equal to the relative price. We will now go through the two key choices using these principles. Before looking at the specific choice examples, let us note that by rationality, in the present context, we include the notion of *perfect foresight*: given that no shocks are occurring, consumers know what prices prevail not only today but also in the future. If there are shocks—and we will study shocks later in the text—rationality is interpreted as *rational expectations*, i.e., knowing the probability distribution for variables in the future.

Consumption vs. saving The relative price between consumption at t and $t+1$ is the *real interest rate*: it is the amount of goods at $t+1$ that a consumer can buy for one unit of the good at t. We will denote the gross real interest rate between t and $t+1$ R_{t+1} here. The marginal rate of substitution between the goods can be obtained by defining an indifference curve relating to these two goods. Thus, write $u(c_t) + \beta u(c_{t+1}) = \bar{u}$, take total differentials, i.e., $u'(c_t)dc_t + \beta u'(c_{t+1})dc_{t+1} = 0$, and then solve for $-dc_{t+1}/dc_t$. Setting the resulting expression equal to the gross real interest rate, we obtain

$$\frac{u'(c_t)}{\beta u'(c_{t+1})} = R_{t+1}.$$

This equation, which equivalently can be written

$$u'(c_t) = \beta u'(c_{t+1})R_{t+1},$$

is commonly referred to as the *Euler equation* and it is a central element of macroeconomic theory. It says that an optimizing consumer sets the marginal utility loss of saving one consumption unit for tomorrow (the left-hand side) equal to the gain tomorrow in consumption

[18] Try to verify this by showing that, if the income allocated to the two goods is increased and consumers can choose between the goods freely, given a fixed relative price, both consumption levels will rise.

terms (the right-hand side), that is, R_{t+1} (the return on the savings) times the marginal utility of each unit tomorrow, $\beta u'(c_{t+1})$.

We argued above that a constant saving rate will imply convergence toward a constant level of capital relative to technology, i.e., $\hat{k}_t$ becomes constant—this is implied by Solow's analysis, which we will elaborate more on later. In particular, a constant saving rate is associated with aggregate consumption growing at a constant rate. We also saw that balanced growth requires a constant real interest rate. So the question now is whether individuals, when faced with a constant interest rate, would choose a consumption path that grows at a constant rate, despite the desires to smooth consumption over time. The question boils down to whether the Euler equation could hold for constantly growing consumption, and we now address this question.

Let us begin with the answer: there is a sharp characterization saying that the utility function u is consistent with exact balanced growth if and only if it is a power function. It is easy to verify the "if" part: $u'(c_t)/u'((1+\gamma)c_t)$ becomes constant if $u(c)$ is a power function, and that constant contains the growth rate. Hence, under a power utility function a constant interest rate will lead the consumer to choose a constant consumption growth rate. What that growth rate is precisely depends on the interest rate, on β, and on the curvature of u, but not on how wealthy the consumer is. The precise class of functions is captured by

$$u(c) = \frac{c^{1-\sigma} - 1}{1 - \sigma} \tag{2.5}$$

with $\sigma > 0$ and $\sigma \neq 1$; the case where σ approaches 1 yields $u(c) = \log c$.[19] The "only if" result is harder to prove, but within reach; the proof is in the appendix to Chapter 4.

The above discussion has been carried out heuristically and in terms of simply selecting two adjacent time periods, without reference to how many periods the consumer lives in total. This discussion also means that the arguments above hold whether in dynastic or overlapping-generation economies, or some combination of these, so that the restrictions placed on preferences in order to be consistent with balanced growth hold rather generally.

Note that the utility-function characterization came from a requirement of exact balanced growth. One could imagine growth paths that are asymptotically balanced and still match our historical data. In the limit, however, the underlying utility functions would then look like our u defined in equation (2.5).[20]

Let us summarize: in macroeconomic modeling, where consumption-saving choices are viewed to come from optimizing consumers, the utility function employed in almost all applications is a power function. This choice is made because we want our frameworks to account for the basic historical facts. For utility functions that are not in this class, we would therefore, for example, see very different saving rates 100 years ago than today; typically, saving rates would go to zero or to one over time as the economy grows, and balanced growth as observed in the data would not be possible.

[19] To understand the $\sigma = 1$ case, take the limit as σ goes to 1 but use l'Hôpital's rule. One obtains

$$\lim_{\sigma \to 1} \frac{d(c^{1-\sigma} - 1)/d\sigma}{d(1-\sigma)/d\sigma} = \lim_{\sigma \to 1} \frac{-c^{1-\sigma} \log c}{-1} = \log c.$$

[20] An example is $u(c + 0.01)$, where u satisfies (2.5).

Labor vs. leisure Turning to labor supply, the idea is to allow households to choose how much to work. But do people choose how much they work? In many countries, work hours are regulated, and your own specific employer may not offer you much choice. From the historical, macroeconomic perspective, however, there is no doubt that labor supply is a choice. First, we have seen that in the longer run, work hours per adult have fallen appreciably. Second, looking across countries, a similar pattern emerges (one we will revisit later): households in rich countries work significantly less than do households in poor countries. These patterns reflect differences (over time and across space) that involve both the intensive margin, i.e., how many hours each individual works when she works, and the extensive margin, i.e., whether a given individual works at all and the fraction of her lifetime the individual works positive hours. We thus see differences across time and space as reflecting choice and, for example, labor-market regulations stipulating a 40-hour workweek should be seen as an outcome reflecting people's choice at the time these regulations were decided upon.[21] In sum, and especially with our long-run perspective, we consider the choice of how much labor to supply as a very natural one.

In more concrete terms, and focusing on the intensive margin only, the period utility function has to allow the agent to explicitly value leisure. Assume that the time endowment is 1 and the consumer chooses between working hours ℓ and leisure l: $\ell + l = 1$. We can choose to have leisure as an argument: u would depend on (c, l), or equivalently, $(c, 1-\ell)$. Below, we write the utility function as $u(c, 1-\ell)$ to represent this dependence.

Now we need to insist on balanced growth in consumption jointly with a constant labor supply in the long run. We therefore need the condition $u_2(c_t, 1-\ell_t)/u_1(c_t, 1-\ell_t) = w_t$, where $u_i(\cdot,\cdot)$ represents the partial derivative with respect to ith argument, to be met at all points in time on a balanced growth path where, as shown above, c and w grow at the same rate.

Let us begin with the requirement that work hours are constant along a balanced path, as motivated by the postwar U.S. data. Then, just like in the above case, there is a sharp characterization of what preferences are consistent with an exact balanced-growth path. In particular, they are consistent with an exact balanced-growth path if and only if the utility function is of the form $u(c, 1-\ell) = ((cv(1-\ell))^{1-\sigma} - 1)/(1-\sigma)$, where $v(l)$ is strictly increasing (in leisure $l = 1-\ell$) and such that $cv(l)$ is strictly quasiconcave. It is again straightforward to show the "if" part—it is a matter of looking at the Euler equation and the first-order condition for the hours choice jointly and verifying that balanced growth is consistent with the derived equations—but, as before, more demanding to show the "only if" part.

A second possibility is that we would like our theory to be consistent also with the longer-run data and across countries with very different standards of living. This is actually possible, since it also seems—though it was perhaps not clear from the earlier graphs—that as output has grown at a roughly constant rate, hours worked have declined at a roughly constant rate. The rate of hours decline is only, say, a third of a percent per year, but over time these small changes accumulate and become visible. It turns out, moreover, that an outcome with exact balanced growth where hours fall at a constant rate is possible as an outcome within our

[21] Of course, in countries that are not democratic, one could imagine that work hours are dictated. But even then people's preferences would likely play a role.

framework if and only if the utility function satisfies $u(c, 1-\ell) = (c^{1-\sigma} g(c^{\frac{\nu}{1-\nu}}\ell) - 1)/(1-\sigma)$, $\nu > 0$, and $g(\cdot)$ is a decreasing function. With technology and output growth proceeding at a net rate γ, we then have hours grow at $(1+\gamma)^{-\nu}$ and consumption at $(1+\gamma)^{1-\nu}$. Wages now grow slightly faster than output, and the labor share remains a constant share of output.[22]

The case just mentioned collapses to the one described earlier with $\nu = 0$. With $\nu = 0$, as wages grow along a balanced path, the substitution effect (making it more beneficial to work at higher wages) exactly cancels the income effect (making it more attractive to choose more leisure as income rises); if $\nu > 0$, the income effect is slightly stronger. Thus, the interpretation is that at lower levels of income, people work more because consumption is more important to them.

In conclusion, we have now arrived at a utility-function specification that is (the only one) consistent with choosing a constant saving rate and constant (or constantly declining, at a low rate) labor supply in the long run. The precise population structure and the length of households' lives can, however, satisfy a variety of assumptions and our main two applications below will be the representative-agent dynasty and the simplest overlapping-generations model.

2.2 The rest of the text

In the next five chapters, we will go over the main macroeconomic tools: (i) the Solow growth model (Chapter 3); (ii) dynamic optimization (Chapter 4); (iii) dynamic equilibrium theory (Chapter 5); (iv) welfare (Chapter 6); (v) uncertainty (Chapter 7); and (vi) empirical methods (Chapter 8). The methods presented in these chapters are core material that will then be used and applied over and over. The methods will also make the material discussed in the present chapter more precise; for example, convergence in the Solow model will be discussed in detail, the maximization problems of consumers will be fully described and a convergence theorem under optimal saving will also be provided, a dynamic competitive equilibrium will be defined and characterized in its application to the growth model, the welfare properties of such equilibria will be discussed in some detail, and it will be made clear under what assumptions the methods and results extend straightforwardly to the case of uncertainty.

The two methodological chapters, Chapters 9 and 10, describe some basic mathematical and computational tools not described in other chapters.

We now very briefly discuss some of the key issues and contents in the applied chapters that then follow in the second part of our text.

The applied issues: Chapters 11–25

Chapter 11: consumption This chapter looks more in detail at how key consumer choices are made on the individual level. It makes clear that the very simplest consumption-saving model delivers predictions for marginal propensities to consume—these are a key component of macroeconomic propagation mechanisms—that are hard to square with micro

[22] Constantly declining, or growing, hours worked do not pose a problem from the perspective of the production side.

data: they are too low. Imperfect insurance against individual shocks to labor-market outcomes, an a priori plausible addition to the basic model, will change this and bring the model closer in line with the data. At this point, heterogeneity will become an important element of the analysis and the connection between macroeconomics and inequality will appear for the first time.

Chapter 12: labor supply As just discussed in the present chapter, the choice of labor supply can be seen as an extension to basic consumption theory. Allowing for both an extensive and an intensive margin of choice is of particular relevance, the former referring to whether or not to work at all and the latter to how many hours to work, conditional on working. Given this foundation, it is of significant interest to understand how labor choice varies with wages: the wage elasticity of labor supply. There are different notions of elasticities and they are useful in different contexts, with a particular distinction between how hours vary over time in response to wage changes and how hours vary across countries. Another key distinction that will be discussed is individual vs. aggregate labor supply.

Chapter 13: growth Now the wide disparity of incomes across countries will be addressed. What explains why many countries remain poor, and more generally how do the distributions of GDP, consumption, and other key variables across all countries in the world evolve over time? What is the role of technological progress, and what determines it? Human capital accumulation and its role in the growth process are also discussed. Relative to the material in the earlier chapters, this chapter introduces some additional features and, at the same time, flesh out quantitative predictions and compare them with data. At the end of the chapter, there will be an attempt at providing quantitative answers to the core growth questions.

Chapter 14: real business cycles Already in Chapters 3 and 7 there are some glimpses into how the neoclassical model fares when the economy is hit by random shocks. The so-called *real business cycle* (RBC) model, as laid out by Kydland and Prescott in their 1982 paper, proposed that macroeconomic fluctuations are indeed to an important extent a result of unforeseen changes (mostly increases in) technology. In addition, and even more importantly, their paper introduced quantitative, microeconomics-based macroeconomic theory as a tool. Their idea was to formulate a stochastic, dynamic general-equilibrium (DSGE) model, solve it numerically for parameter values that were plausible given micro data and long-run facts, and then use it to address macroeconomic phenomena such as fluctuations and the effects of policy changes. Virtually all analyses of macroeconomic fluctuations since have adopted their approach, though many of the frameworks that came later departed in various ways from Kydland and Prescott's basic RBC model. In particular, many other shocks were proposed, money was introduced, and numerous frictions to how markets operate were included. This chapter thus begins by discussing how to filter the raw data in order to focus on the business-cycle frequencies and describes the key facts: how macroeconomic aggregate fluctuate and correlate. It then describes the core model, if nothing else because many of the later models treat it as a benchmark. What is the jury's verdict on the role of technology shocks in accounting for business-cycle fluctuations? This question will be

addressed in the later chapters.

Chapter 15: government So far, very little has been said about the government. The U.S. government is sizable, and many other governments are much larger still as a share of total expenditures or employment. Governments spend resources and make transfers aimed at redistribution, e.g., from rich to poor and across age groups. The chapter first describes and discusses the key facts pertaining to government variables. Second, the chapter uses our basic theory to examine, given some specified objectives, how different government financing schemes compare, both positively and normatively. Taxation, for example, tends to involve distortions, so some effort will be devoted to understanding its effects. Is it important that the government runs a balanced budget, and does debt management even matter? As a part of this effort, the chapter will give us an introduction to how one can formulate optimal policy problems aiming to maximize consumer welfare while being restricted to the use of distortionary taxes.

Chapter 16: asset prices The previous chapters will have studied many intertemporal issues, including borrowing and lending, but the analysis of asset markets, and the determination of asset prices in particular, is important in its own right. The chapter will thus describe the key facts—for example, asset prices fluctuate "wildly" and risky assets pay a much higher return on average than do riskless assets such as U.S. Treasury bonds—and then proceed to analyze these facts through the lens of our basic theories. A core framework is the so-called consumption Capital Asset Pricing Model (CAPM), which derives asset prices and their stochastic features in relation to explicit household choices over its stochastic consumption path.

Chapter 17: money A very particular asset is fiat money. "Fiat" refers to the fact that money is intrinsically without value and, nowadays, the value of money is in no sense backed by any real objects (as it was historically in many economies: its value was backed by gold). This raises the question of why it has value at all. From this perspective, inflation means that money loses value. The chapter will go over basic data on inflation and basic theories of how the value of money can be determined, all in the absence of price-setting frictions. It will briefly touch on the determination of exchange rates too; although the value of money in terms of real goods in any given economy does not necessarily fluctuate much, exchange rates do and the chapter will briefly address this volatility. The chapter, finally, will provide a bridge into the next chapter by explaining how money is introduced in the so-called New-Keynesian model of business cycles, where price-setting frictions are central, as well as a discussion of the interdependence of fiscal and monetary policy.

Chapter 18: nominal frictions and business cycles Prices and wages appear to move sluggishly on in the micro data. The chapter will begin by documenting some key facts on this and also review some evidence suggesting that monetary policy can have real effects because prices and wages are "sticky." The New-Keynesian model will then be introduced here. This model has become a workhorse for central banks around the world. It builds on the RBC model but adds nominal frictions: costs associated with changing prices and wages. The

extension to the RBC model involves introducing long-lived firms with market power: these firms set prices knowing that prices will be costly to change in the future. The framework also has a description of how the central bank behaves; in particular it introduces a notion of monetary "policy shocks," as an additional source of macroeconomic fluctuations. The chapter will discuss the evidence on the role of monetary policy in accounting for aggregate fluctuations.

Chapter 19: frictional credit markets By many economists, the Great Depression is viewed to have in part been caused by frictions in the credit market, i.e., impediments to borrowing for firms. Similarly, the 2007–2009 Great Recession is also considered to have had its roots in financial-market malfunctioning. The chapter will begin by documenting some correlations that suggest that financial frictions might be important. It will then show how such frictions can be introduced into the core framework and how macroeconomic propagation sometimes, but not always, changes nature in the presence of such frictions.

Chapter 20: frictional labor markets Often, the rate of unemployement is even used to define the business cycle: it is highly countercyclical—rises in recessions and falls in booms. The chapter begins by reviewing not only the key facts on aggregate unemployment but also on individuals' movements in and out of jobs over time. It then introduces the most common framework for analyzing unemployment: the search and matching model. It begins by looking at worker search and then introduces a full general-equilibrium model with matching frictions as in Pissarides (1985a). The resulting model is then confronted with data and the so-called Shimer puzzle is introduced and discussed. Finally, the chapter shows how the Pissarides model can be extended so as to incorporate capital accumulation and, thus, as such can be seen as an important extension of our basic macroeconomic framework.

Chapter 21: inequality In this chapter, we discuss inequality between households. The chapter views inequality as interesting in its own right and, hence, reviews both data and theory. The focus is broad, thus covering labor-market inequalities (wages, earnings, and hours) as well as inequality in consumption and in wealth, and for each variable of interest it surveys the main theories. The discussion, again, aims to be quantitative, i.e., the theories are evaluated based on how much of the observed inequality they can plausibly account for. The chapter also asks how inequality might matter for macroeconomic aggregates. We have already touched on one way in which it could: to the extent a model with significant inequality generates a marginal propensity to consume that is higher on average, it will alter many of the model's predictions. This is an active research area in macroeconomics; the HANK model—a Heterogeneous-Agent New-Keynesian setting—in particular has already had significant impact in applied monetary policy contexts.

Chapter 22: heterogeneous firms The introduction of the workhorse model is based on an aggregate production function. Clearly, this is an abstraction, at the same time as it hopefully offers a good approximation to the properties of a more realistic framework with a multitude of firms. This chapter examines this issue, both by looking at data on firms and by constructing models with firm heterogeneity. Like the chapter on household heterogeneity, it

discusses how firm heterogeneity suggests new mechanisms and thus add insight into how the macroeconomy works. Two channels are studied in particular. One involving misallocation of input factors across firms when there are frictions. The other makes specific assumptions about firm size and discusses granularity: a notion of extreme firm inequality where some very large firms can be relevant to the whole economy. The chapter also briefly touches on markups and the degree of competition.

Chapter 23: international macroeconomics Many readers of this textbook will perhaps not primarily feel at home in the "large, closed economy" version of our macroeconomic theories. There are, in fact, even strong arguments to suggest that the U.S. economy of today is much more dependent on the rest of the global economy than it used to be and, therefore, issues of trade, exchange rates, and international borrowing and lending ought to take a more central place than it does in many textbooks. The present chapter thus tries to make amends. In particular, it builds toward an up-to-date international business cycle model with monetary and other frictions. The focus is on conceptual issues and mechanisms rather than on a full quantitative model.

Chapter 24: sovereign debt and default risk Yet many other readers may feel that the focus of our text is on the highly developed, world-leading economies such as the U.S. while their interest at least in part is in macroeconomic issues in emerging markets: countries that have opened up to trade and hope to grow rapidly to begin catching up with the leading countries. These countries are argued to have specific vulnerabilities, for example in their ability to borrow in times of severe recessions; indeed sovereign default has been commonly observed, i.e., episodes where debts are not paid back and capital flight occurs. This chapter discusses these issues, again as an extension of our workhorse model: what features need to be added, or changed, to deliver a framework that can be used to study macroeconomic fluctuations in emerging markets?

Chapter 25: sustainability The final chapter should concern all readers, independent of country of origin, as it deals with global topics that have risen to the top of the political agenda virtually everywhere: areas where human economic activity causes environmental problems, such as climate change. The specific question of climate change concerns macroeconomics, as macroeconomic activity, at least historically, is closely tied to carbon dioxide emissions—as a byproduct of using fossil fuels for energy generation. The chapter mostly focuses on climate change and thus goes through the necessary basic natural-science background and the way in which climate and economics interact. A simple "integrated assessment model" is developed and used to study how policy can be used to address the issue. A brief discussion of natural resource use is also included.

Chapter 3

The Solow model

One of the long-run economic facts presented in Chapter 2 is the stability of the capital-output ratio K_t/Y_t (where Y_t represents aggregate output at period t and K_t represents aggregate capital at period t) over time: capital is roughly three times annual GDP. Earlier contributions of growth theory, such as the so-called Harrod-Domar model, considered this stability as representing a technological property and incorporated it as one of the assumptions in the model. The capital stock, traditionally divided into structures and equipment but nowadays also containing some intangible components (e.g., software), is one of the important production inputs, but it is of course not the only one. It is possible to produce products in a very capital-intensive way, but clearly there is a choice and using labor—different people's time, skills, and effort—is the most obvious alternative, or complement. Given the many possibilities in which production process can be set up, it is therefore not obvious why, at the macro level, K_t/Y_t is almost constant over time. As argued in Chapter 2, this was Solow's starting point and he managed to resolve the tension between the stable aggregate ratio and the intuitive notion that capital and labor are quite substitutable by a sequence of insights that led both to the construction of a framework for studying macroeconomic dynamics and for measurement of technological change. The purpose of the present chapter is thus to detail the Solow model: the basic assumptions underlying it and their implications.

A central element in the Solow model is the aggregate production function. In the aggregate production function, there are three economic variables that can affect the growth of GDP: technology A_t, capital K_t, and labor L_t. The Solow model focuses on the endogenous accumulation of capital K_t. We will see that K_t not only reacts to the saving rate but also to A_t and L_t. After solving the model, which will deliver a stable capital-output ratio, we will focus on two main takeaways: (i) the fundamental source of long-run growth in per capita income is the growth in A_t; (ii) if all parameter values are common, different economies converge to the same (both in terms of level and growth rate) income per capita in the long run.

3.1 The basic model

We start our exposition using the simplest version of the model, where there is neither technological progress nor any growth in the size of the population or the skills of workers. The centerpiece of the Solow model is the aggregate production function

$$Y_t = F(K_t, L_t).$$

Note that we can interpret Y_t as the GDP in this economy. We make the following assumptions for the function $F(K, L)$:

1. $F(K, L)$ is strictly increasing in both K and L.
2. $F(K, L)$ is strictly quasiconcave in (K, L) (it has strictly convex isoquants).
3. $F(K, L)$ exhibits constant returns to scale in (K, L): when K and L change to cK and cL, with any $c > 0$, $F(K, L)$ becomes $cF(K, L)$.
4. $F(0, L) = 0$.
5. $\lim_{K\to 0} F_1(K, L) = \infty$, where $F_1(K, L) \equiv \partial F(K, L)/\partial K$.
6. $\lim_{K\to\infty} F_1(K, L) = 0$.

Assumptions 5 and 6 are often called Inada conditions and are stronger than we need but these assumptions simplify the exposition.[1]

In the basic model, we assume that the population is constant and that hours worked per worker is constant, so that L_t is constant. We normalize both population and hours per capita to 1; therefore, the only variable input for production is K_t, and because of this normalization, we can write $F(K_t, 1) = F(k_t, 1)$, where we remind the reader that lower-case letters are per-capita measures. Let us use $f(k_t)$ to denote $F(k_t, 1)$. Then the production function can be expressed

$$y_t = f(k_t). \tag{3.1}$$

The second important piece of the Solow model is the equation that describes the evolution of the capital stock:

$$k_{t+1} = i_t + (1 - \delta)k_t, \tag{3.2}$$

where i_t is investment in period t. The existing capital stock loses value, δk_t, while being used, where $\delta \in (0, 1)$ is capital's depreciation rate.

These two centerpieces are connected through individual behavior. First, the goods supply y_t is equal to the demand for goods, $c_t + i_t$:

$$y_t = c_t + i_t. \tag{3.3}$$

Here, c_t is consumption and i_t is investment (both per capita). Note that here we implicitly assume (as in the large part of the following chapters) that goods are homogeneous and can

[1] We also assume that F is twice continuously differentiable. This means that first-order conditions to maximization problems involving F can be differentiated and then generate continuous functions.

be used for both consumption and investment. Because output y_t is also the total income for consumers, it is either consumed or saved. Therefore, we know that total saving has to equal total investment. In an open economy—where there is trade—this does not necessarily hold. Finally, in this chapter section, we can interpret both c and i as including government consumption and investment, respectively.

In the Solow model, instead of explicitly modeling the consumption-saving decisions of consumers, the consumers are assumed to mechanically save a constant fraction of their income, so that investment is given by

$$i_t = sy_t, \tag{3.4}$$

where $s \in (0, 1)$ is the constant saving rate. This behavioral assumption is relaxed and replaced by consumers' optimizing consumption-saving behavior in Chapter 4.

Inserting equation (3.4) into equation (3.2) and using equation (3.1) yields

$$k_{t+1} = (1 - \delta)k_t + sf(k_t). \tag{3.5}$$

This difference equation expresses the dynamics of the capital stock k_t over time. This is *the fundamental equation of the Solow model*. Note that the only endogenous variable on the right-hand side of the fundamental equation is k_t. Therefore, the next period's stock of capital k_{t+1} can be determined only with the knowledge of the current capital stock k_t, given values of exogenous objects: the scalars δ and s and the function f. Note also that, starting from a given k_0, once we obtain the series of $\{k_{t+1}\}_{t=0}^{\infty}$ from the fundamental equation (3.5), the time series $y_t = f(k_t)$, $c_t = (1 - s)y_t$, and $i_t = sy_t$ can readily be obtained.

3.1.1 Steady state and dynamics

To analyze the difference equation (3.5), we first consider a special situation where k_t is constant over time. Call this situation the *steady state* and denote it with an upper bar: $k_t = \bar{k}$ for all t. From the fundamental equation (3.5), the steady-state capital stock can be determined by the solution of the equation

$$\bar{k} = (1 - \delta)\bar{k} + sf(\bar{k}).$$

This equation implies $\delta\bar{k} = sf(\bar{k})$. It is straightforward to verify that, under the assumptions for the aggregate production function in the previous section, a strictly positive value of $\bar{k}$ that solves this equation always exists and is unique. Graphically, plot the left- and right-hand sides of the equation $\delta\bar{k} = sf(\bar{k})$; the left-hand side is a straight line through the origin with a positive slope and the right-hand side, which also starts in the origin, is strictly increasing and strictly concave, with a slope of infinity at 0 and one that approaches 0 as $\bar{k} \to \infty$. Clearly, we see that an intersection exists and is unique. The Inada conditions are used here to guarantee the existence of $\bar{k}$. In the context of the basic model and as pointed out above, the Inada conditions are stronger than necessary: they can be replaced by weaker versions $\lim_{k\to 0} F_1(k, 1) > \delta/s$ and $\lim_{k\to\infty} F_1(k, 1) < \delta/s$.

A particularly useful production function is the Cobb-Douglas production function: $F(K, L) = K^{\alpha}L^{1-\alpha}$, so that $f(k) = k^{\alpha}$, where $\alpha \in (0, 1)$. This production function satisfies all assumptions we need, including the Inada conditions. With Cobb-Douglas production, $\bar{k}$ can be

solved for analytically:

$$\bar{k} = \left(\frac{s}{\delta}\right)^{\frac{1}{1-\alpha}}. \tag{3.6}$$

From this expression, we can see that the capital stock in the steady state is increasing in the saving rate s and decreasing in the depreciation rate δ. Because aggregate output (GDP) is $\bar{y} = f(\bar{k})$, $\bar{y}$ is also increasing in s and decreasing in δ.

Now, let us use a diagram to analyze the dynamics of k_t when $k_0 > 0$ is not at the steady-state level. Figure 3.1 plots the equation (3.5) with the 45-degree line (that is, representing $k_{t+1} = k_t$). In the figure, the intersection of (3.5) and the 45-degree line represents the steady-state $\bar{k}$.

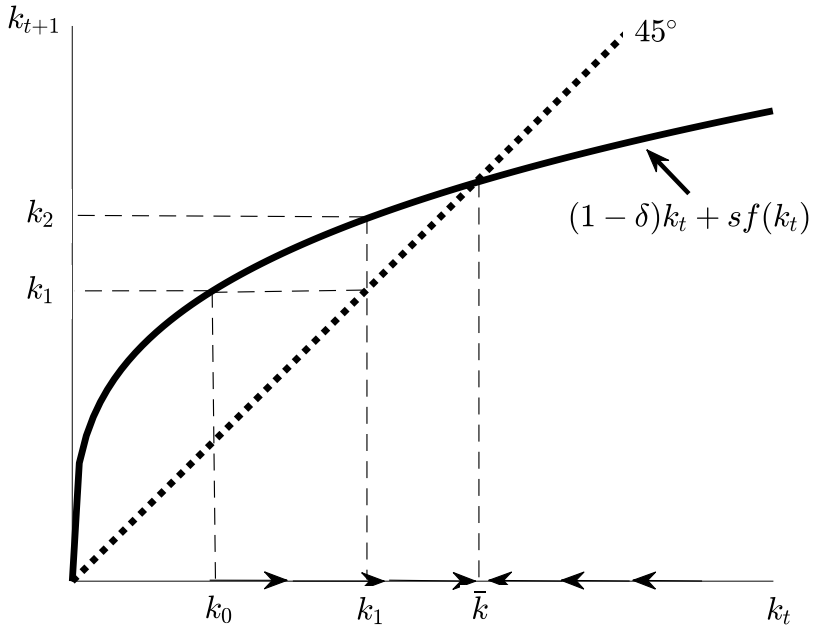


Figure 3.1: Dynamics in the Solow model

In the figure, when we start from a given k_0 on the horizontal axis, we can obtain k_1 on the vertical axis by using the (3.5) curve. By placing this k_1 back on the horizontal axis and using the (3.5) curve again, we obtain k_2, and so on. This procedure yields the full time series of the capital stock: $\{k_{t+1}\}_{t=0}^{\infty}$. One can easily verify the dynamics of k_t exhibits a global and monotonic convergence to $\bar{k}$, regardless of the initial value k_0. That is, whatever the starting point is, the time series of k_t gradually approaches $\bar{k}$ over time.

The figure works very well as a graphical argument, but how would a mathematical proof be put together? Suppose that $k_t < \bar{k}$. It is then straightforward to show that (i) $k_{t+1} > k_t$ (because $sf(k_t) > \delta k_t$ when $k_t < \bar{k}$) and also that (ii) $k_{t+1} < \bar{k}$ (because $k_t < \bar{k}$ and the right-hand side of (3.5) is increasing in k_t). Repeating this procedure, we can see the sequence $\{k_t, k_{t+1}, k_{t+2}, ...\}$ is monotone and bounded by $[k_t, \bar{k}]$. From the Monotone Convergence Theorem, the sequence has a limit. The limit has to be $\bar{k}$, as the limit is unique under the conditions given.

Intuitively, the convergence occurs because the aggregate production function $f(k_t)$ has

decreasing returns to capital (Assumption 2 above). Equation (3.2), rewritten in terms of the change in the capital stock,

$$k_{t+1} - k_t = i_t - \delta k_t,$$

reveals two forces that go in opposite directions: (gross) investment and depreciation. When the total capital stock is small, output per unit of capital is large, and the constant saving rate then implies that a large (gross) investment is made relative to the existing capital stock. This process enables the aggregate capital stock to increase. As k_t increases, output per unit of capital becomes smaller due to the decreasing returns property, and when k_t is very large enough, the gross investment cannot cover total deprecation, δk_t. Thus, the investment force is stronger when k_t is small and the depreciation force is stronger when k_t is large. This relationship is perhaps even clearer if we write the above equation in terms of the growth rate:

$$\frac{k_{t+1} - k_t}{k_t} = \frac{sf(k_t)}{k_t} - \delta,$$

where we have replaced $i_t = sf(k_t)$. The assumptions $f''(k_t) < 0$ and $f(0) = 0$ imply that $f(k_t)/k_t$ is decreasing in k_t, generating the negative relationship between the investment force of pushing up the capital stock and the level of the capital stock. In fact, when saving behavior (as captured by s here) is modeled explicitly as a choice, s can counteract this force toward convergence, but it turns out not to be strong enough to overturn the convergence result. This issue will be discussed in detail in the next chapter.

Let us go back to our motivating fact: the stability of k_t/y_t over time. In the basic model here, k_t/y_t is of course constant in steady state, as are all the variables. Below, however, we will see that, even in a growing economy where k_t and y_t keep increasing over time, the economy settles to a situation where k_t/y_t is constant over time.

In the Cobb-Douglas case above, the steady-state k/y ratio can be solved out as

$$\frac{\bar{k}}{\bar{y}} = \frac{s}{\delta}.$$

Clearly, in the long run k_t/y_t is larger when s is larger and when δ is smaller.

Other kinds of dynamics

The growth model can, in principle, generate very rich (and complex!) dynamics if its neoclassical feature is not present, i.e., if the production function is not strictly concave in capital. There are applications in the economics literature that, in reduced form, have such non-neoclassical features, and we now briefly illustrate how they can work.

Endogenous growth Consider the situation where $F_1(k, 1)$ is uniformly above δ/s. Figure 3.2 below draws such an example. In this case, the steady state with $\bar{k} > 0$ does not exist, and k_t keeps growing larger over time. That is, there is unbounded growth "by itself": growth in *endogenous*. This concept will be discussed more in Chapter 13 below but in a richer model where other production inputs can also be accumulated. Here, given that one

expects decreasing returns to each input—such as capital—it is hard to take this case very seriously.

In the special (and illustrative) case where F_1 is a constant—as illustrated in the graph—we can think of output as linear in capital: $y_t = Ak_t$, with no role for labor (make $\alpha = 1$ in the Cobb-Douglas setting).[2] Given that labor commands about two thirds of the income from production, this setup does not seem empirically plausible. In a setup with endogenous growth such as this, two identical countries starting out with different capital stocks will be forever different; the gap between them, in percentage terms, will stay constant.

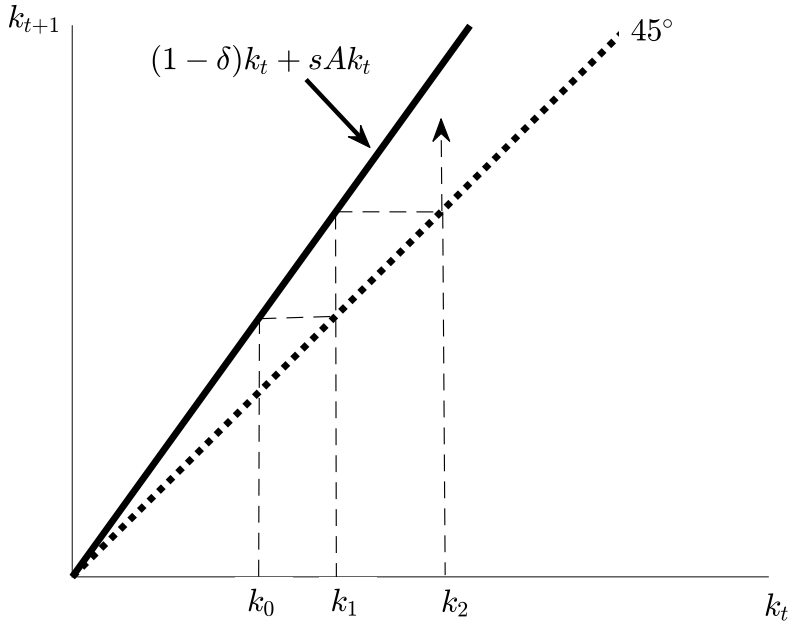


Figure 3.2: Endogenous growth in the Solow model

Poverty traps

Suppose that the production function is not globally concave in k: it has a middle section that is convex. This could be true if there are some regions of k with increasing returns, say, as a result of large infrastructure investments—the building of transportation networks. In such a case, the right-hand side of (3.5) will not be concave, and it may cross the 45-degree line multiple times, as illustrated in Figure 3.3 below.

[2] One can imagine a role for labor if $Y_t = AK_t + BL$, which is CRS, but here labor would not matter asymptotically if $A > \delta/s$.

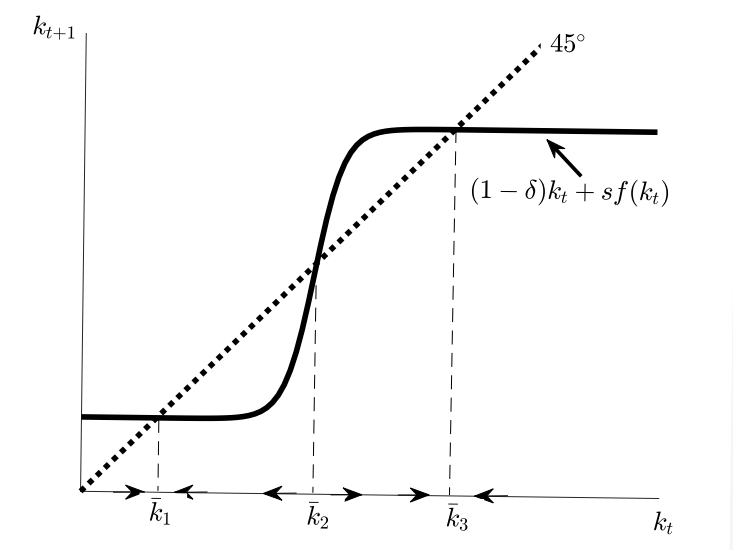


Figure 3.3: Poverty traps in the Solow model

Clearly, in this case there are multiple steady states and at least one of the steady states will then not be "stable": k_t will not converge to that steady state even when k_0 starts very close to it (a small perturbation away from the steady state leads further away from it). When there are multiple steady states, an economy can get stuck in the steady state with a low $\bar{k}$ (and thus a low GDP) when it starts from a low k_0. This situation is often called the *poverty trap*. The gap between a poor country with a small k_0 and a rich country with a large k_0 may never close in this setting. In Figure 3.3, there are three steady states, $\bar{k}_1$, $\bar{k}_2$, and $\bar{k}_3$. Of these three, $\bar{k}_1$ and $\bar{k}_3$ are stable steady states. When the economy starts from a very low k_0, the economy converges to $\bar{k}_1$ and gets trapped in it. To escape from the trap, the capital stock would have to be pushed up to a larger level than $\bar{k}_2$, from which it would converge to $\bar{k}_3$. One way to achieve this movement is to (temporarily) encourage very high saving. If the saving rate is raised permanently so that the $(1-\delta)k_t + sf(k_t)$ curve moves up sufficiently, the steady states $\bar{k}_1$ and $\bar{k}_2$ will disappear and the economy converges globally to $\bar{k}_3$. The growth/development literature does not appear to have identified sufficiently large increasing returns leading to results of the kind described here, but it is an interesting possibility.

Non-monotonic dynamics and chaos An even more radical departure from the neoclassical setting is if $f(k)$ declines in k at high levels of k. Conceptually, if a bakery has no ovens, ovens have high marginal productivity, and as more ovens are added, the marginal productivity declines, and it will become negative once there are so many ovens in the bakery that there is neither space for bakers nor for the dough. This possibility is more esoteric in a macroeconomic context but let us nevertheless study it briefly. So when $f(k)$ decreases steeply enough, the right-hand side of (3.5) will become decreasing in k_t. Illustrating this graphically, we will see that convergence, if convergence is at all possible, will not be monotonic.[a] In fact, k_t can exhibit forever oscillating (or even

chaotic) dynamics.[b] An example of is drawn in Figure 3.4.

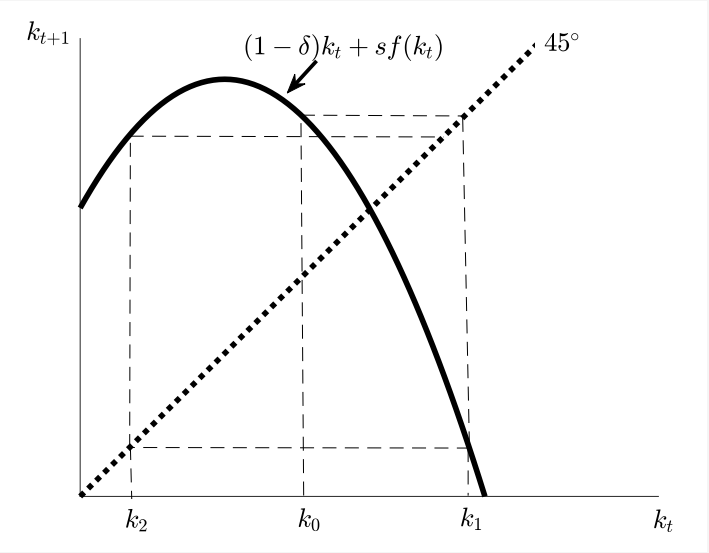


Figure 3.4: Complex dynamics in the Solow model

[a]Locally stable dynamics will occur if the slope at the steady state is less than 1 in absolute value.

[b]Chaos is a mathematical term; it involves great sensitivity to initial conditions and forever non-monotone behavior that never settles down to a repeated pattern (a repeated pattern could be a two-cycle): it looks "random."

3.2 The growing economy

Now we extend the model to the situation where A_t and L_t grow over time. In the previous section, the long-run outcome was the steady state where there is no growth. This extension is necessary for addressing the facts related to the growth issues described in Chapter 2.

We assume that the aggregate production function takes the form of

$$Y_t = F(K_t, A_t L_t).$$

There are two changes from the basic model: first, we allow the labor input (population times hours worked per person) L_t to grow over time. Second, and more importantly, we allow for technological progress. In this production function, the variable representing the technology level, A_t, is multiplied by labor input L_t. Technological progress thus takes a form of improving the labor input, and $A_t L_t$ is often referred to as the total number of *efficiency units of labor* (or *effective labor*). This form of technological progress is labor-augmenting; it was introduced in the previous chapter.[3] As was also asserted there, Uzawa (1961) proved that labor-augmenting technical change is the only form of technical progress that is consistent with exact balanced growth, that is, the growth path where aggregate variables such as output and capital grow at a constant rate. Uzawa's theorem is formally stated and proved in Appendix 3.A.

We assume that the (net) growth rate of A_t is γ and that the growth rate of L_t is n.[4]

[3]This form of technical progress is sometimes also called Harrod-neutral.

[4]n has two origins: a growing population and changes in hours worked per person, which we saw from Chapter 2 is best characterized by a decline in the longer run.

The same manipulations of equations as in the basic model yield

$$K_{t+1} = (1-\delta)K_t + sF(K_t, A_tL_t).$$

Dividing both sides by A_tL_t, we obtain

$$\frac{K_{t+1}}{A_tL_t} = (1-\delta)\frac{K_t}{A_tL_t} + s\frac{F(K_t, A_tL_t)}{A_tL_t}.$$

Let us define a new variable $\tilde{k}_t$ by

$$\tilde{k}_t \equiv \frac{K_t}{A_tL_t}.$$

Then, because the above equation can be rewritten as

$$\frac{A_{t+1}}{A_t}\frac{L_{t+1}}{L_t}\frac{K_{t+1}}{A_{t+1}L_{t+1}} = (1-\delta)\frac{K_t}{A_tL_t} + sF\left(\frac{K_t}{A_tL_t}, 1\right),$$

we obtain

$$(1+\gamma)(1+n)\tilde{k}_{t+1} = (1-\delta)\tilde{k}_t + sf(\tilde{k}_t). \tag{3.7}$$

Here, $f(\tilde{k}_t) \equiv F(\tilde{k}_t, 1)$ as in the basic model. After the capital stock K_t is normalized by A_tL_t, we thus obtain a very similar difference equation as the fundamental equation (3.5) in the basic model. Once we characterize the dynamics of $\tilde{k}_t$, we can "untransform" it into the core macro variables Y_t, K_t, and C_t.

3.2.1 Balanced growth and dynamics

The characterization of the fundamental equation (3.7) follows similar steps as for the basic model. The concept corresponding to the steady state in the basic model is the *balanced growth path* (some researchers still prefer to use the name "steady state" for the balanced growth path, because the normalized variables are "steady" also in this case). Along the balanced growth path, the normalized capital stock $\tilde{k}_t$ is constant, and typical economic variables, such as Y_t, K_t, and C_t, grow at a constant rate. Once again, we use a notation with upper bar: $\tilde{k}_{t+1} = \tilde{k}_t = \bar{\tilde{k}}$. The value of $\bar{\tilde{k}}$ along the balanced-growth path solves

$$(1+\gamma)(1+n)\bar{\tilde{k}} = (1-\delta)\bar{\tilde{k}} + sf(\bar{\tilde{k}}).$$

With a Cobb-Douglas production function we can obtain a closed-form solution:

$$\bar{\tilde{k}} = \left(\frac{s}{(1+\gamma)(1+n)+\delta-1}\right)^{\frac{1}{1-\alpha}}. \tag{3.8}$$

The dynamic property of the model can be analyzed similarly to the basic model. As in the basic model, starting from any $\tilde{k}_0$, $\{\tilde{k}_{t+1}\}_{t=0}^{\infty}$ converges monotonically to $\bar{\tilde{k}}$. In (3.8), the rate of technological progress γ and the population growth rate n work similarly to depreciation: maintaining a level of $k_t = K_t/(A_tL_t)$ is harder as A and L grow faster; i.e., each unit of untransformed capital needs to grow faster, as if making up for depreciation.

In this framework, we can analyze how various economic variables grow over time. For example, suppose L_t is simply population size (assuming that all citizens work one unit) and then consider income per capita, defined as $y_t \equiv Y_t/L_t$. Because $Y_t/(A_tL_t) = f(\tilde{k}_t)$, in the long run, $Y_t/(A_tL_t)$ converges to $f(\bar{\tilde{k}})$. Therefore, in the long run, income per capita converges to

$$y_t = f(\bar{\tilde{k}})A_t$$

and the growth rate of y_t in the long run is

$$\frac{y_{t+1} - y_t}{y_t} = \frac{f(\bar{\tilde{k}})A_{t+1} - f(\bar{\tilde{k}})A_t}{f(\bar{\tilde{k}})A_t} = \frac{A_{t+1} - A_t}{A_t} = \gamma.$$

The growth in technology A_t is essential in sustaining long-run growth in per capita income. Surprisingly, no other parameters affect the long-run growth of per capita income. For example, encouraging saving (an increase in s) does not affect the long-run growth rate of per capita income in the economy. Note that this result does not mean that the change in s does not have any effect on economic outcome: it has an effect on the *level* of per capita income, rather than the growth rate. It also has an effect on the growth rate in the short run (when the economy is not yet on the balanced-growth path).

In the short run, $\tilde{k}_t$ changes over time and its movement has an effect on the economic outcome. For example, the growth rate per capita income is now

$$\frac{y_{t+1} - y_t}{y_t} = \frac{f(\tilde{k}_{t+1})A_{t+1} - f(\tilde{k}_t)A_t}{f(\tilde{k}_t)A_t} = \frac{f(\tilde{k}_{t+1})}{f(\tilde{k}_t)}(1+\gamma) - 1.$$

From the fundamental equation (3.7), we know that when $\tilde{k}_t < \bar{\tilde{k}}$, $\tilde{k}_t$ increases over time, that is, $\tilde{k}_{t+1} > \tilde{k}_t$. Therefore, in this case, $f(\tilde{k}_{t+1})/f(\tilde{k}_t) > 1$ and the growth rate of y_t in the short run is larger than γ. Similarly, when $\tilde{k}_t > \bar{\tilde{k}}$, the growth rate of y_t is smaller than γ. In other words, the Solow model predicts that income per capita of a poor country grows faster than at rate γ and that income per capita of a rich country grows slower than at rate γ in the short run. This difference in growth rates is another representation of the convergence prediction of the Solow model.

3.3 Stylized facts and the Solow model

The model with growth, presented above, can match various stylized facts of economic growth. First, going back to our motivating fact on K_t/Y_t, because $Y_t/(A_tL_t) = f(\tilde{k}_t)$ and $K_t/(A_tL_t) = \tilde{k}_t$ are both constant along the balanced growth path, $K_t/Y_t = \tilde{k}_t/f(\tilde{k}_t)$ is also constant in the long run. Once again, the Solow model can replicate the constant K_t/Y_t in the data.

The first fact in Chapter 2 was the steady growth of the GDP per capita. As we have seen above, the GDP per capita grows at the rate γ along the balanced growth path (towards which the economy converges from any starting point). This fact, therefore, is consistent with the Solow model with technological progress.

Another stylized fact is that the return to physical capital has been nearly constant. Here, we need to first compute the return to physical capital in the model. Suppose that firms maximize profit under competitive markets:

$$\max_{K_t, A_t L_t} F(K_t, A_t L_t) - r_t K_t - w_t A_t L_t. \tag{3.9}$$

Here, output is taken as the numéraire and its price is set at one. Therefore, $F(K_t, A_t L_t)$ is the revenue and $r_t K_t + w_t A_t L_t$ is the cost. Let us assume, for simplicity, that the capital stock is owned by the consumers and rented to the firms with the rental rate r_t. Thus, r_t represents the return to capital. The other component of the cost is the wage payment: w_t is the wage per efficiency unit of labor. From the first-order condition for K_t, the return to physical capital is equal to the marginal product of capital:

$$r_t = F_1(K_t, A_t L_t).$$

Differentiating both sides of the equation $f(K_t/A_t L_t) = F(K_t, A_t L_t)/(A_t L_t)$ with respect to K_t, we obtain that

$$r_t = f'(\tilde{k}_t).$$

Along the balanced growth path, therefore, r_t is constant because the right-hand side is constant at $f'(\bar{\tilde{k}})$.

Another prominent fact is the stability of the labor share and the capital share. The capital share is equal to $r_t K_t / Y_t$, and it can readily be seen that it is constant, because we have already seen that both r_t and K_t/Y_t are constant along the balanced-growth path. The labor share is $w_t A_t L_t / Y_t$. From the first-order condition of (3.9), the wage is equal to the marginal product of labor:

$$w_t = F_2(K_t, A_t L_t).$$

When the production function has constant returns to scale, it is homogeneous of degree one.[5] Then it follows that production becomes

$$Y_t = K_t F_1(K_t, A_t L_t) + A_t L_t F_2(K_t, A_t L_t).$$

Dividing both sides by Y_t, we obtain that the labor share is one minus the capital share. Therefore, the labor share is also constant when the capital share is constant.

We can also confirm the stability of the labor share by direct calculation. As for the case of r_t, it can be shown, by differentiating $f(K_t/A_t L_t) = F(K_t, A_t L_t)/(A_t L_t)$ with respect to $A_t L_t$, that

$$w_t = f(\tilde{k}_t) - \tilde{k}_t f'(\tilde{k}_t).$$

It can readily be seen that w_t is constant when $\tilde{k}_t$ is constant at $\bar{\tilde{k}}$. Because $A_t L_t / Y_t = 1/f(\tilde{k}_t)$, it is also constant along the balanced growth path. Therefore, $w_t A_t L_t / Y_t$ is also constant. Note that, for a Cobb-Douglas production function $Y_t = K_t^{\alpha}(A_t L_t)^{1-\alpha}$, the capital share is α and the labor share is $1-\alpha$ regardless of the values of K_t and $A_t L_t$, and therefore the factor shares are constant even outside the balanced-growth path.

[5] Recall that a function $f(x)$ is homogeneous of degree r (is $H(r)$) when $f(sx) = s^r f(x)$ for all (s, x); here x is a vector and r and s are scalars. If $r = 1$, it then follows, using differentiation with respect to s and each element of x, that $f(x) = \sum_i (\partial f/\partial x_i) x_i$ for all x.

3.4 Convergence

We have already seen that, in the Solow model, the economy monotonically converges to the steady state (or balanced growth path). Here, we look at this convergence property more in detail and take a quick look at the data.

3.4.1 Local properties: the speed of convergence

In the basic model, where we again set $L_t = 1$ and use the per-capita notation k_t, the fundamental equation (3.5) can be approximated around the steady-state as

$$\Delta k_{t+1} = \left(1 - \delta + sf'(\bar{k})\right) \Delta k_t, \tag{3.10}$$

where Δk_t represents the deviation of k_t from its steady-state value, that is, $k_t - \bar{k}$, when the deviation is small.[6] When the production function is in the Cobb-Douglas form, using the steady-state solution (3.6),

$$\Delta k_{t+1} = (1 - \delta(1 - \alpha))\Delta k_t$$

holds. Replacing Δk_t by $k_t - \bar{k}$ and dividing both sides by $\bar{k}$, (3.10) can be expressed as

$$\frac{k_{t+1} - \bar{k}}{\bar{k}} = (1 - \lambda)\frac{k_t - \bar{k}}{\bar{k}},$$

where $\lambda \equiv \delta - sf'(\bar{k})$ represents the *convergence speed*. A large value of λ implies that Δk_{t+1} becomes smaller (in absolute value) more quickly, implying a faster convergence. This is illustrated in Figure 3.5 below: a higher λ represents a flatter slope at the steady state and "more steps until you reach steady state."

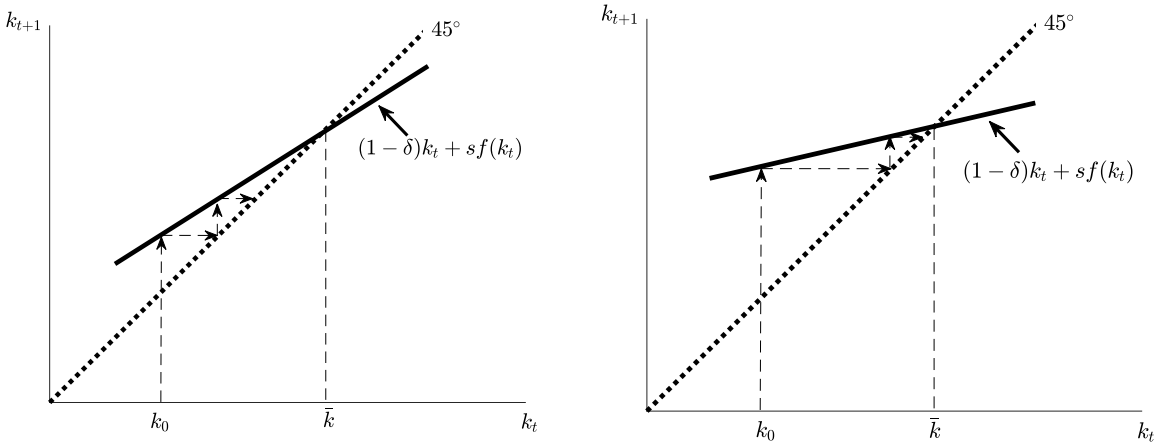


Figure 3.5: Slow and fast convergence

In the Cobb-Douglas case,

$$\lambda = \delta(1 - \alpha) \tag{3.11}$$

[6] This is obtained from a first-order Taylor approximation of the expression around $\bar{k}$.

holds, and thus the convergence is faster when α is small and δ is large. Note that s does not affect the convergence speed in this case. The convergence speed, in general, is affected by how k_t affects k_{t+1}. In an extreme case, if k_t has no effect on k_{t+1} (a flat line), convergence is immediate. The parameter s has two opposing forces to this mechanism. For a given k_t, a large s implies a larger impact of k_t on k_{t+1}. However, the steady-state value of capital is larger when s is larger, and thus the marginal product of capital at the steady state, $f'(\bar{k})$, is smaller when s is larger, implying a smaller impact of k_t on k_{t+1}. These two opposing forces exactly offset each other when the production function is in the Cobb-Douglas form.[7]

In the case with growth, the fundamental equation (3.7) can be approximated by

$$(1+\gamma)(1+n)\Delta\tilde{k}_{t+1} = \left(1-\delta+sf'(\bar{\tilde{k}})\right)\Delta\tilde{k}_t. \tag{3.12}$$

When the production function is in the Cobb-Douglas form, using the steady-state solution (3.8),

$$\Delta\tilde{k}_{t+1} = \left(\alpha + \frac{(1-\alpha)(1-\delta)}{(1+\gamma)(1+n)}\right)\Delta\tilde{k}_t$$

holds. The equation (3.12) can be rewritten as

$$\frac{\tilde{k}_{t+1}-\bar{\tilde{k}}}{\bar{\tilde{k}}} = (1-\lambda)\frac{\tilde{k}_t-\bar{\tilde{k}}}{\bar{\tilde{k}}},$$

where the convergence speed is now given by $\lambda = 1-(1-\delta+sf'(\bar{\tilde{k}}))/(1+\gamma)(1+n)$. In the Cobb-Douglas case we obtain

$$\lambda = (1-\alpha)\left(1-\frac{1-\delta}{(1+\gamma)(1+n)}\right). \tag{3.13}$$

3.4.2 Cross-country data

Is convergence observed in the data? Recall that the convergence prediction implies that a country that starts with a smaller per-capita GDP experiences faster subsequent growth. Figure 3.6 plots this relationship across countries. The data is taken from the Penn World Table 10.0 (https://www.rug.nl/ggdc/productivity/pwt/). The horizontal axis is per-capita real GDP in 1960, which we take as the starting point. The vertical axis is the subsequent growth rate (annualized using geometric averages) in per-capita real GDP from 1960 to 2019.

One can immediately see that there is no systematic tendency for initially poor countries to grow faster. The fact that there is no tendency for countries to converge, however, does not imply a rejection of the Solow model. In fact, the Solow model does not predict that different countries will always converge to the same balanced growth path (a phenomenon called "unconditional convergence" or "absolute convergence"). Rather, it predicts that countries converge if they share the same parameter values ("conditional convergence"). We

[7] The Cobb-Douglas function is very convenient because it often simplifies the algebra and leads to simple expressions. This simplicity, however, can be deceiving as we see here: the functional form often makes fundamental forces going in opposite direction cancel. That is, under the surface there may be very strong forces but, as if by magic, the Cobb-Douglas form makes them invisible.

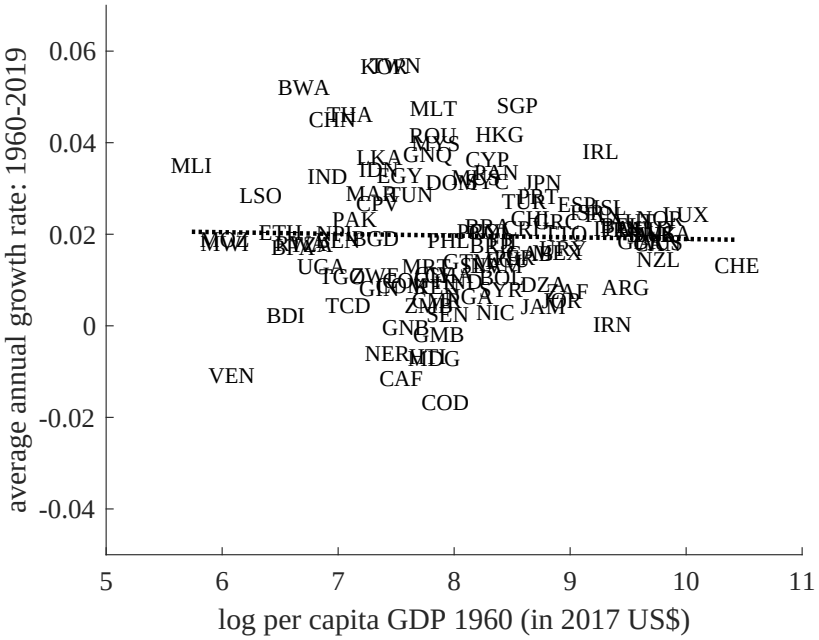


Figure 3.6: All countries, 1960–2019

Source: Penn World Tables 10.0. The GDP variable used is RGDPNA.

know, for example, that saving rates differ widely across countries and, at least over shorter time horizons, it is reasonable to think that the growth rates of A_t also differ.

To examine conditional convergence, a useful exercise is to look at the same kind of graph restricted to a smaller, and more similar, set of countries. Figure 3.7 thus plots the same data as Figure 3.6, but only for the original members of the Organisation for Economic Co-operation and Development (OECD). OECD was formed by high-income countries that share relatively similar economic and political institutions, and we can expect the underlying parameters in the Solow model to be relatively similar among these countries.

For this set of countries, we observe a clear tendency for convergence: poor countries in 1960 on average experience faster subsequent growth. Barro and Sala-i Martin (1995) (Chapter 12) conduct a similar exercise across regions within countries, treating each region as a different "country." For U.S. states, Japanese prefectures, and European regions, they find a clear tendency of convergence, supporting the prediction of the Solow model.

In a recent paper, Kremer, Willis, and You (2022) show that in recent years, the data actually show a tendency for unconditional convergence. Figure 3.8 below repeats the same exercise as in Figure 3.6 for the same set of countries, but setting the initial date to 2000. We can see that some convergence (negative correlation) is observed in the recent years. The authors argue that this tendency arose because some of the underlying factors that affect growth (the factors that likely affect the growth rate of A_t), such as policies, institutions, and human capital have become more similar across countries in recent years. In Chapter 13, we will discuss this and many related issues in greater detail.

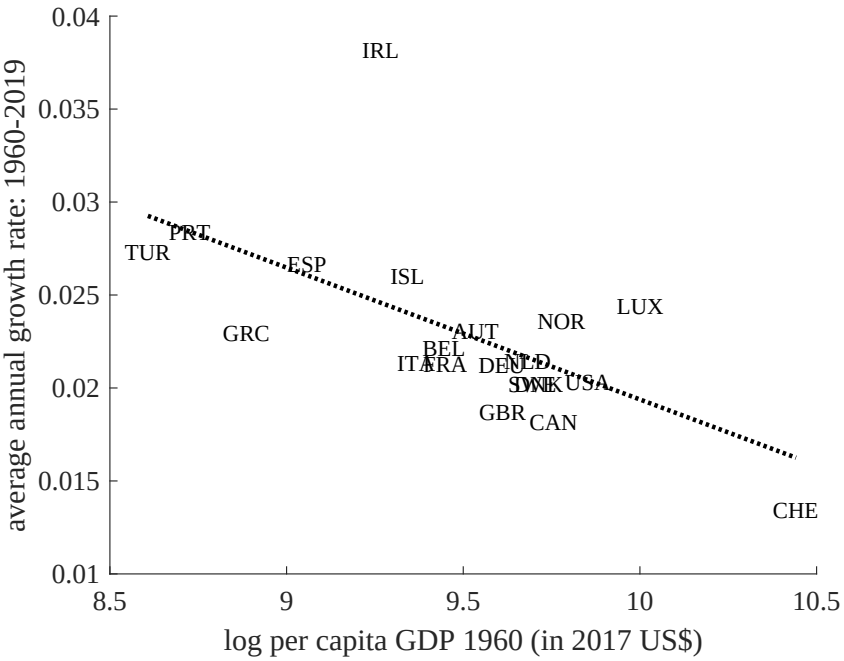


Figure 3.7: OECD countries, 1960–2019

Source: Penn World Tables 10.0. The GDP variable used is RGDPNA.

3.4.3 Quantitative use of the Solow model

What are the *quantitative* predictions of the model for convergence? To answer this question, we need to assign functional forms and specific parameter values. This procedure will give us a numerical value for λ in (3.13). Once λ is computed, one can of course also conduct counterfactual experiments by simulating a hypothetical situation using the quantitative model.

If we are interested in the model's quantitative predictions for the speed of convergence, one way to proceed would be to simply see if it is possible to choose parameters so as to hit the "observed λ" (e.g., as measured by the slope in Figure 3.7). More generally, one could specify a full stochastic model, say, with explicit shocks to variables (such as A_t) and estimate the resulting structure against the data we just looked at. Clearly, we could generate a good fit in this case if we are free to choose parameters; for example, given any production function, we could match the λ by an appropriate choice of δ. This choice, however, may not be consistent with what we know about depreciation rates from microeconomic data. More generally, we would like our model's different components (functional forms and parameter values) to be selected to be in line with microeconomic studies (and perhaps aggregate data too). This way, the quantitative evaluation is disciplined. As briefly discussed in Chapter 1 above, a way forward is *calibration*.[8] The procedure consists of two distinct steps, each guided by data. First, we assign a parameterized functional forms to the unknown functions

[8]We describe calibration, and compare it to other methods, in much more detail in Chapter 8.

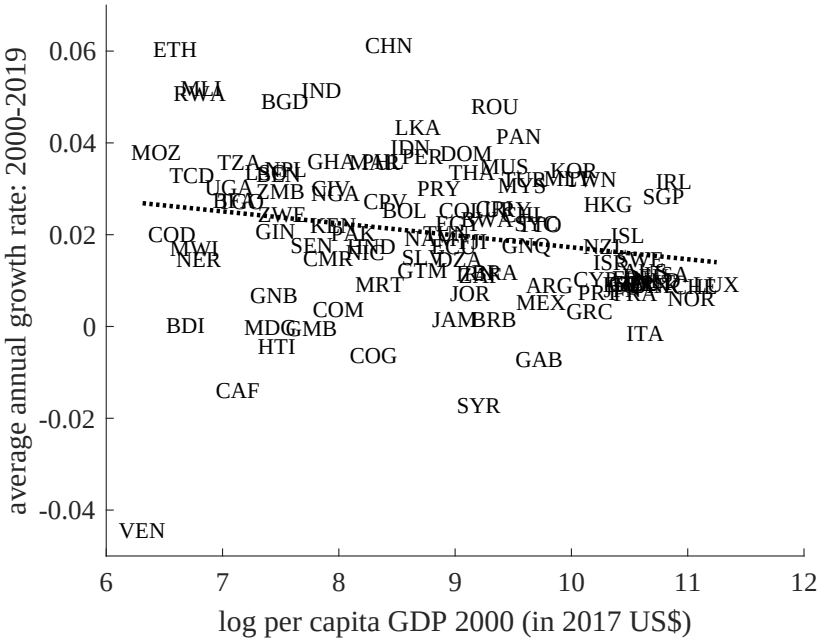


Figure 3.8: All countries, 2000-2019

Source: Penn World Tables 10.0. The GDP variable used is RGDPNA.

in the model. In our case, the production function F corresponds to the unknown function. Second, we assign specific values to the parameters. Because we use various moments (such as means and variances) in the data to assign parameter values, this second step bears close resemblance to estimation by the method of moments.

Before starting the calibration, we have to decide on the length of the time period. Here, we are chiefly interested in movements in aggregates that occur over the medium run and thus set one period to be one year. For the first step, we choose the Cobb-Douglas form, used above, for the production function: $F(K_t, A_t L_t) = K_t^{\alpha}(A_t L_t)^{1-\alpha}$. As we have seen, this functional form yields constant factor shares, which is consistent with the rather striking data on an absence of major movements in the shares. A Cobb-Douglas function is special, however, in that it has the property that the substitution elasticity between inputs (capital and labor) is equal to one, so we need to make sure that it is consistent with studies of production functions, at least at a high level of aggregation.[9] These studies rarely suggest major departures from 1.

In the second step, we need to assign values to the parameters α, δ, γ, n, and s. Here, we have implicitly assumed that the production function has constant returns to scale, i.e., that the sum of the exponents on capital and labor is one, so that it suffices to choose the α. Before discussing the parameter selections, note that for the convergence speed λ in (3.13), the information on s is not necessary. Below we will therefore skip assigning a value to s.

[9]The substitution elasticity is given by the percentage change in the ratio of the inputs when the ratio of their prices change by one percent, i.e., $-d\log(K/L)/d\log(r/w)$. Using the firm's first-order conditions, we can see that this expression must equal 1 for a Cobb-Douglas production function.

Calibration usually draws on multiple data sources. Overall, there are two methods of assigning the parameter values based on data. First, if particular parameters have been considered as important objects for investigation elsewhere and we know plausible parameter values from past studies, it is convenient to directly assign the parameter values accordingly. Second, we can choose data moments that involve several parameters and assign parameter values such that these moments, when generated by our theory, line up with observed moments. The first method can be considered a special case of the second method, because the parameter values from past studies have to come from certain data moments used in these studies. From this perspective, an alternative interpretation of the calibration procedure is an implementation of the method of moments with multiple data sets.

We set the value of α from national income accounting. As we have seen from the previous section, α corresponds to the capital share. From Figure 2.12 in Chapter 2, $\alpha = 1/3$ is a good approximation. We can set the value of γ from the long-run growth rate of per capita income in advanced countries. The population growth rate n can be measured directly from the data. Barro and Sala-i Martin (2004) use $\gamma = 0.02$ and $n = 0.01$ as a benchmark at the annual frequency. When $\tilde{k}_{t+1} = \tilde{k}_t$ (along the balanced growth path), the fundamental equation (3.7) can be rewritten as

$$(1+\gamma)(1+n) = 1 - \delta + \frac{I_t}{K_t},$$

where we used $sf(\tilde{k}_t)/\tilde{k}_t = I_t/K_t$. The investment-capital ratio in the U.S. economy is about 0.076 (Cooley and Prescott, 1995) at an annual frequency. Given the above values of γ and n, this equation implies $\delta = 0.046$.[10] Clearly, here, one could alternatively have used depreciation rates directly from data on depreciation (by capital type, or from aggregate data on depreciation rates) and then this equation would have implied a value for the average size of I/K on a balanced growth path. Given the accounting practices and the fact that capital remains at roughly three times annual GDP as time passes, on average investment precisely will have to make up for depreciation (taking population and technology growth into account), so either way the number for δ ends up around 0.05.

With these values, we find that $\lambda = 0.049$. The empirical counterpart of λ is about 0.015 to 0.03 (Barro and Sala-i-Martin 2004, p,59), and therefore the Solow model over-predicts the convergence speed. The model value of λ is about 0.02 if α is raised to 0.73. An argument for a larger value of α is that a part of labor income is the return to human capital, which can be accumulated in a similar manner as physical capital, and thus some part of labor income should be included in the capital share. Relatedly, one can view A_t as an accumulable factor—after all, technological development is often part of conscious investments into R&D, and hence it too is a capital stock. These issues are returned to in Chapter 13.

Once the model has been assigned specific functional forms and parameter values, we can also conduct quantitative experiments. Suppose, for example, the saving rate s equals 0.1. How would increasing s to 0.2 affect the normalized level of output along the balanced-growth path, $f(\bar{\tilde{k}})$? With a Cobb-Douglas production function, we have already obtained

[10] Given that $I/K = (I/Y)(Y/K)$, we could alternatively have measured I/Y—the saving rate—and Y/K, which we know is about 1/3. Conversely, we can now obtain the implied I/Y as 0.076/0.33, which approximately equals 0.23.

the solution for $\bar{\tilde{k}}$ in (3.8). Inserting our calibrated parameter values, along with $s = 0.1$, we obtain $\bar{\tilde{k}} = 1.20$ and $f(\bar{\tilde{k}}) = \bar{\tilde{k}}^{\alpha} = 1.06$. When s goes up to 0.2, $\bar{\tilde{k}}$ rises to 1.90 and $f(\bar{\tilde{k}}) = 1.24$. Therefore, doubling the saving rate from 10% to 20% increases the normalized level of output by 17%, because $1.24/1.06 = 1.17$. In addition, we could compute the quantitative predictions for how fast output would rise to eventually reach a 17% higher value.

3.5 Business cycles

The usefulness of the Solow model goes much beyond the study of economic growth; due to its ability to account for the broad features of the macroeconomic data over our modern economic history, it constitutes the core of macroeconomic modeling. A prominent illustration of this is the fact that virtually all modern theories of business cycles build on a version of, or elaboration on, the Solow model. Although business cycle theories are detailed later in Chapters 14 and 18, we now briefly review how these theories are linked to the Solow model and exhibit some of the key associated tools, such as impulse response diagrams.

3.5.1 Various theories of business cycles

The studies of business cycles are primarily the analysis of the arrival of shocks to the economy and how the economy reacts to these shocks. The way the economy reacts to the shocks is usually called the "propagation mechanism." Once we introduce uncertainty in Chapter 7, we can treat these "shocks" more precisely. Here, we consider a general (exogenous and deterministic) movement in certain variables as the source of business cycle fluctuations.

Below, first, we extend the basic model in several directions. In particular, we outline how the Solow model can be modified and accommodate various shocks in three different business cycle models. Throughout we conduct the analysis in per-capita terms and, hence, use lower-case letters.

- The first business cycle model is the so-called real business cycle (RBC) model. As will be explained in Chapter 14, the RBC model provides a simple mechanism whereby macroeconomic variables comove, as is clear in the data. The prototypical RBC model considers the following aggregate production function

$$y_t = A_t F(k_t, \ell_t),$$

  where ℓ_t is variable labor input, and considers a shock to A_t (often called the "neutral technology shock"). That is, it assumes that A_t changes over time and the movement of A_t is the source of the business cycle. For example, consider a model where A_t switches around between two values, A_H and A_L, where $A_H > A_L$. When A_t moves to A_H, the economy starts moving towards the corresponding steady state. Then A_t switches to A_L, and the economy now moves towards a different steady-state value. We can interpret these movements as business cycles: movements around some average. Augmented with steady growth we would have movements around the balanced path.

Note that the production function in this section is different from the Harrod-neutral form $F(k_t, A_t\ell_t)$ earlier. Note also that the distinction between Harrod-neutral technological progress and the Hicks-neutral technological progress is not essential when the production function is in the Cobb-Douglas form: the Harrod-neutral production function $k_t^\alpha(A_t\ell_t)^{1-\alpha}$ can be rewritten as $A_t^{1-\alpha}k_t^\alpha\ell_t^{1-\alpha}$, and by defining $\tilde{A}_t \equiv A_t^{1-\alpha}$, the same production can be interpreted as the Hicks-neutral production function $\tilde{A}_t k_t^\alpha \ell_t^{1-\alpha}$.

If we maintain the assumptions of the basic model, we have ℓ_t constant and $i_t/y_t = s$ (and $c_t/y_t = 1 - s$) constant even with the shocks to A_t. These features are at odds with the business cycle data. To accommodate the business cycle facts that (i) ℓ_t comoves positively with the business cycle, (ii) i_t is more volatile than y_t, and (iii) c_t is less volatile than y_t, the basic equation would have to allow for the saving rate and ℓ_t to react to A_t (and possibly to k_t). The economy evolves, therefore, following the difference equation

$$k_{t+1} = (1 - \delta)k_t + s(k_t, A_t)A_t F(k_t, \ell(k_t, A_t)).$$

This is a modified form of the fundamental equation (3.5) of the basic Solow model.

- Next, we consider a different kind of shock. Suppose that the final goods market clearing condition (3.3) is modified to

$$y_t = c_t + i_t/\nu_t,$$

where ν_t moves over time (and often called the "investment-specific technological progress"): when it is high, it is cheaper to produce investment goods. One can think of this equation as reflecting the two-sector structure of the economy: y_t and c_t are measured in consumption goods, and investment goods have a production process that can create i_t units of investment goods by using i_t/ν_t units of consumption goods. The fundamental equation (3.5) can now be modified to

$$k_{t+1} = s\nu_t F(k_t, \ell) + (1 - \delta)k_t;$$

this equation can also be extended to include endogenous s and ℓ as in the case of the neutral technology shock.

- In the third example, we consider a very different model structure: one with "demand shocks." First, assume that c_t is exogenous and that it moves around over time. The movement of c_t serves as the (demand) shock. Suppose, further, that we maintain the assumption that $i_t/c_t = s/(1 - s)$. Because of this assumption, since s is constant, i_t is proportional to c_t and moves along with it. Therefore, the total demand for goods is a function of c_t:

$$c_t + i_t = \frac{1}{1 - s}c_t.$$

When c_t is not sufficiently large, $c_t + i_t$ would be less than the full capacity output $F(k_t, \ell)$ (here we assume again that ℓ is given by labor-force participation: those who want to work). Assume, then, that when there are demand shortages, total output y_t

is determined by the demand side and so that a fraction u_t of the labor force become unemployed (therefore, u_t is the unemployment rate):

$$y_t = \frac{1}{1-s} c_t = F(k_t, \ell(1-u_t)).$$

From the second equality, u_t can be represented as the function of c_t and k_t: $u(c_t, k_t)$. The fundamental equation (3.5) can therefore be modified as

$$k_{t+1} = sF(k_t, \ell(1-u(c_t, k_t))) + (1-\delta)k_t.$$

This framework is very Keynesian in spirit but clearly begs the question of how output can end up below full capacity, and hence be driven by demand. In a well-functioning market, this phenomenon could not occur. This book contains several chapters on frictions that could lead to something like the setting just described, and it then becomes central for policymakers to understand the exact nature of these frictions.[11]

In all three cases above, we can represent k_{t+1} as a function of k_t and a shock (A_t, ν_t, or c_t). This representation allows us to characterize the dynamics of k_t (and other macroeconomic variables, such as y_t, c_t, and i_t) in response to these shocks.

3.5.2 Impulse responses

One method of describing how the economic variables respond to shocks is to draw an impulse-response function. Consider the RBC example above, with a Cobb-Douglas production function, an exogenous saving rate, and fixed labor supply. Suppose that before period 0, the value of A_t is constant at $\bar{A}$. After a sufficiently long time, the value of k_t settles close to the corresponding steady-state value $\bar{k}$. Then suppose that at time 0, A_0 is $(\varepsilon \times 100)\%$ higher, that is, $A_0 = (1+\varepsilon)\bar{A}$. For $t = 1, 2, 3, \ldots$, the value of A_t is $A_t = (1+\rho^t\varepsilon)\bar{A}$, where $\rho \in (0, 1)$.

We can then generate the resulting time-path of k_t, starting from $k_0 = \bar{k}$, with

$$k_{t+1} = sA_t k_t^{\alpha} \ell^{1-\alpha} + (1-\delta)k_t, \tag{3.14}$$

for $t = 0, 1, 2, \ldots$. This time path is called the impulse-response function. The time-path of A_t is drawn in Figure 3.9—the impulse—along with the response of k_t. For the impulse-response function for k_t we use $s = 0.2$, $\delta = 0.046$, and $\alpha = 1/3$. The starting value of A and the value of ℓ are normalized to 1. The initial value of the shock to A, ε, is 1%, and the persistence $\rho = 0.9$.

In the figure, k_t increases from its steady-state value of 9.066 to a maximum value of 9.089 and then gradually goes back to the steady-sate value: a hump shape. Three properties of the graph are worthwhile emphasizing. First, the movement of k_t is much slower than the movement of A_t. The deviation of A_t becomes very close to zero around period 50, whereas the response of k_t is more persistent. Second, the magnitude of the response of k_t is

[11] Chapter 21, in particular, describes a related example where c is a "production externality" and, as such, acts like a demand channel.

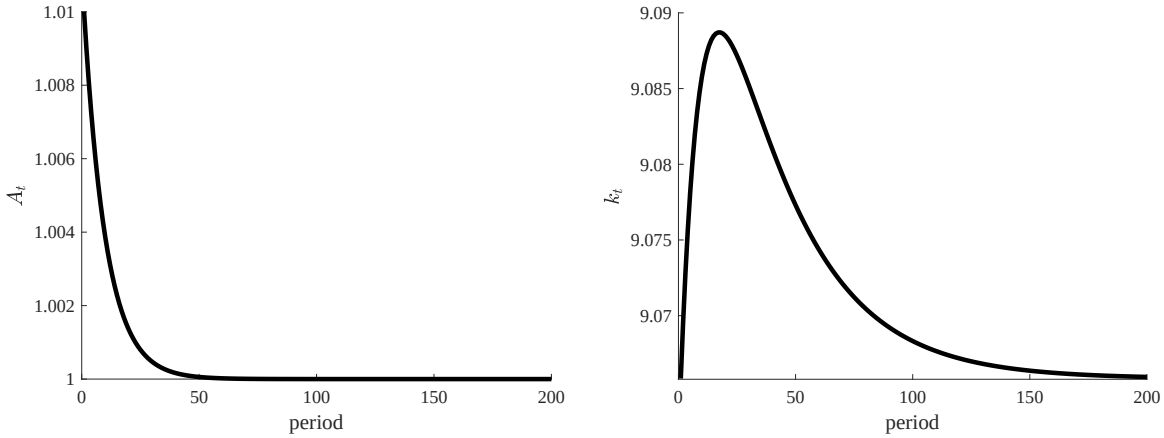


Figure 3.9: Impulse response: how A_t (left) affects k_t (right)

significantly smaller than the impulse: the maximum deviation as a fraction to the steady-state level is $9.089/9.066 - 1 = 0.0025$, that is, 0.25%. This magnitude is much smaller than the initial A_t deviation of 1%. Third, k_t eventually comes back to the steady-state value. The convergence force is always at work when we consider the response to recurrent shocks, as in the business-cycle examples.

Log-linearized impulse responses

Often it is more convenient to approximate the dynamics around the steady state by a log-linarized system. Log-linearization expresses the system in terms of deviation of the logarithms of variables from their corresponding steady-state values. An arbitrary variable x_t can be expressed as

$$x_t = \bar{x}e^{\hat{x}_t}, \tag{3.15}$$

where

$$\hat{x}_t \equiv \log\left(\frac{x_t}{\bar{x}}\right),$$

is the log deviation from the steady-state value $\bar{x}$. Note that because

$$\hat{x}_t \equiv \log\left(\frac{x_t}{\bar{x}}\right) \approx \frac{x_t - \bar{x}}{\bar{x}},$$

where the approximation is the first-order Taylor expansion around $\bar{x}$, $\hat{x}_t$ can be interpreted as the percent deviation from the steady state. Also note that

$$\bar{x}e^{\hat{x}_t} \approx \bar{x}(1 + \hat{x}_t) \tag{3.16}$$

from the first-order Taylor approximation of the expression.

Consider the impulse-response experiment of the system (3.14). Using the transformation

(3.15),

$$\bar{k}e^{\hat{k}_{t+1}} = s\bar{A}e^{\hat{A}_t}(\bar{k}e^{\hat{k}_t})^{\alpha}\ell^{1-\alpha} + (1-\delta)\bar{k}e^{\hat{k}_t}$$

holds. Using (3.16) and the fact that

$$\bar{k} = s\bar{A}(\bar{k})^{\alpha}\ell^{1-\alpha} + (1-\delta)\bar{k} \tag{3.17}$$

holds from the definition of $\bar{k}$, we obtain

$$\bar{k}\hat{k}_{t+1} = s\bar{A}(\bar{k})^{\alpha}\ell^{1-\alpha}(\hat{A}_t + \alpha\hat{k}_t) + (1-\delta)\bar{k}\hat{k}_t.$$

Because equation (3.17) implies $\delta\bar{k} = s\bar{A}(\bar{k})^{\alpha}\ell^{1-\alpha}$, this equation can be rewritten as

$$\hat{k}_{t+1} = (1-\delta(1-\alpha))\hat{k}_t + \delta\hat{A}_t. \tag{3.18}$$

Let us again use $\lambda \equiv \delta(1-\alpha)$, which is the notation for the convergence speed in (3.11). In fact, the log-linearization procedure here is essentially the same as the procedure for obtaining the percentage deviation in the convergence section, in the sense that both are applying the first-order Taylor approximation to (3.14).

The above impulse-response experiment corresponds to setting $\hat{A}_0 = \varepsilon$ and $\hat{A}_t = \rho^t\varepsilon$ for $t = 1, 2, 3, \ldots$. Solving (3.18) with this specification yields

$$\hat{k}_{t+1} = \sum_{\tau=0}^{t}\rho^{t-\tau}(1-\lambda)^{\tau}\delta\varepsilon = \rho^t\frac{1-\left(\frac{1-\lambda}{\rho}\right)^{t+1}}{1-\frac{1-\lambda}{\rho}}\delta\varepsilon.$$

Quantitatively, the log-linear approximation performs well in our calibrated model. The maximum error (in the units of K_t) of the approximation is about 0.00001. If we were to plot the log-linear solution and the nonlinear solution in the same figure, the difference would not be visible.

Applying the above log-linearization procedure to the production function $y_t = A_t k_t^{\alpha}\ell^{1-\alpha}$, the log-deviation of output is

$$\hat{y}_t = \hat{A}_t + \alpha\hat{k}_t = \rho^t\delta\varepsilon + \alpha\rho^{t-1}\frac{1-\left(\frac{1-\lambda}{\rho}\right)^{t}}{1-\frac{1-\lambda}{\rho}}\delta\varepsilon = \rho^t\left(1 + \frac{\alpha}{\rho}\frac{1-\left(\frac{1-\lambda}{\rho}\right)^{t}}{1-\frac{1-\lambda}{\rho}}\right)\delta\varepsilon.$$

for $t = 1, 2, \ldots$ (for $t = 0$, $\hat{y}_0 = \hat{A}_0 = \varepsilon$). This equation explicitly describes how y_t moves over the cycle, in response to the realization of the shock ε. Log-linearization allows us to derive this explicit expression.

An alternative to the log-linear approximation is a linear approximation in *levels*. Let the production function be $A_t F(k_t, \ell_t)$. Assuming that $\ell_t = 1$ for any t and defining $f(k_t) \equiv F(k_t, 1)$, the fundamental equation (3.5) (with the modification in the production function) can be written as

$$k_{t+1} = g(k_t, A_t),$$

where $g(k_t, A_t) \equiv (1-\delta)k_t + sA_t f(k_t)$. Using the notation we employed earlier,

$$\Delta k_{t+1} = g_k\Delta k_t + g_A\Delta A_t,$$

where g_k and g_A are the partial derivatives of $g(k, A)$ with respect to k and A, respectively (evaluated at the steady state), and $\Delta k_t = k_t - \bar{k}$ and $\Delta A_t = A_t - \bar{A}$. When $f(k_t) = k_t^{\alpha}$, we obtain $g_k = 1 - \delta(1-\alpha)$ and $g_A = s^{\frac{1}{1-\alpha}}\delta^{-\frac{\alpha}{1-\alpha}}A^{\frac{\alpha}{1-\alpha}}$.

Chapter 4

Dynamic optimization

In the Solow model presented in Chapter 3, the evolution of the capital stock, along with technological change, is the driving force of output growth. In the model, because agents are assumed to save a constant proportion s of income, saving and, therefore, investment decisions are exogenously given. Cass (1965) and Koopmans (1963) developed the first optimizing models of growth by adding *microeconomic foundations*, i.e., by describing how saving comes about as a result of (rational) choice. These foundations were based on the inter-temporal trade-off between consumption today and consumption in the future, also known as the consumption-saving model. Augmenting the Solow model to incorporate endogenous investment decisions by individuals gave rise to the "optimizing neoclassical growth model," which is nowadays at the core of modern macroeconomic theory. This model delivers the same long-run implications as the original Solow model, but because individual utility is explicit, the model can also be used for analyzing normative issues (individual welfare outcomes). In addition, because the decisions are forward-looking, we can obtain richer policy implications even on the positive side. With the introduction of methods dealing with stochastic model elements ("shocks"), it also serves as the foundation of the real business cycle theory, which we will discuss in more detail in Chapter 14.

In this chapter, we introduce a simple dynamic optimization model and discuss the main characteristics of this class of models. A key objective is to describe how to characterize solutions to dynamic optimization problems, and to introduce discrete dynamic programming methods. We work with a representative agent who faces a dynamic trade-off, meaning that sacrifices today yield gains in the future. For example, in the standard consumption-saving model, the agent can save today (reducing current consumption) in exchange for increasing consumption possibilities in the future. In such a model, the saving rate s is endogenously determined. When extending this model to incorporate production, saving and investment are optimally determined. Because saving depends on income in the neoclassical growth model, the level of capital and the productivity level will now affect the saving rate.

A key underlying assumption, first introduced by Milton Friedman in 1957, is that individuals are *forward-looking*. This means that they do not make decisions based just on their current income (as in the textbook Keynesian model) but also considering their expected future income. In other words, consumption and saving decisions depend on expectations. Subsequent research on consumption also incorporated the idea of *rational expectations*, which states that individuals use all the information available to them at each point in time

to make the best possible forecast about the future. While this assumption may seem extreme, it is arguably a good first approximation to the average behavior of individuals for important decisions in their lives. Deviations from rational expectations are studied in behavioral macroeconomics, a subfield of our subject that has important contributions but we regard as second-year material and therefore do not cover in this book. In the present chapter, we assume that agents are fully rational and forward-looking. Because their future income is taken into account when making dynamic decisions, it is important to determine the time horizon of their decision-making process. There are two common approaches followed in the literature: (i) agents live a finite number of periods, or (ii) agents live forever. The latter is interpreted as a *dynastic* structure in which individuals alive today care about the welfare of their descendants, as discussed later in this chapter. Because the infinite-horizon models require more mathematical sophistication, we start with a finite-horizon model. We also discuss two alternative ways of solving dynamic optimization problems: using sequential methods and using recursive methods. Sequential methods involve maximizing over sequences. Recursive methods, also labeled dynamic programming methods, involve functional equations and characterize choices as *decision rules*. We begin with sequential methods and then move to recursive methods.

4.1 A dynamic optimization problem

Economic decisions are made by agents. These could be: (i) individuals (or households) deciding how much of a good or service to consume, (ii) firms choosing how much to produce, or (iii) a government deciding on policy. In this chapter, we focus mainly on the decisions of individuals. The problem of the firm is described in detail in Chapter 5, whereas the government policy decisions are deferred to Chapter 15.

Agents live for T periods, and this time horizon can, in principle, be finite or infinite. With reference to the next chapter, which deals with market economies inhabited by many consumers, we assume that consumers are all identical, so we now study a **representative agent**. This agent makes decisions over sequences of **allocations** in order to maximize a lifetime objective. Allocations are quantities of goods or services, such as consumption, hours worked, investment, etc. The typical dynamic optimization problem in sequential form studied in macroeconomics takes the following form:

$$\max_{\{y_t, x_{t+1}\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t \hat{\mathcal{F}}(y_t) \tag{P1}$$

subject to

$$x_{t+1} = h(x_t, y_t)$$

and

$$x_{t+1} \in \Gamma(x_t).$$

In this problem, $\sum_{t=0}^{T} \beta^t \hat{\mathcal{F}}(y_t)$ represents the **objective function**. The summation between 0 and T indicates that decisions must be made for each and all of those time periods. The function $\hat{\mathcal{F}}$ is the instantaneous objective, representing, for example, a per-period utility function (for individuals) or profit function (for firms). The constant β is

referred to as the stationary discounting weights: it is our discount factor; $1/\beta - 1$ is the discount rate. They are called stationary because the ratio between the weights of any two different dates $t = i$ and $t = j > i$ only depends on the number of periods elapsed between i and j, and not on the values of i or j. In other words, $\beta^{t+k}/\beta^t = \beta^k$.

The sequence $\{y_t\}_{t=0}^T$ represents the choice variables, sometimes referred to as **control variables**. Examples of these are consumption, leisure, saving, and investment levels at each point in time. The sequence $\{x_t\}_{t=0}^T$ represents the **state variables**. Examples are the stocks of capital, debt, or housing. States and controls are related through the constraint $x_{t+1} = h(x_t, y_t)$. For example, h can reflect a budget constraint or a production technology. The value of the initial state x_0 is exogenously given. To differentiate states from controls, notice that if an agent chooses y_0 in period 0, the value of the state next period, x_1, is automatically determined from $x_1 = h(x_0, y_0)$ because x_0 is exogenous.[1] In period 1, the choice of the control y_1 determines the state in the following period, x_2, and so on. In other words, choosing the control variable optimally at t determines the value of the state variable at $t + 1$. Finally, $\Gamma(x_t)$ represents the feasible set which, given the value of the current state x_t, restricts the values that x_{t+1} can take; we will be specific on standard forms Γ can take later.

For dynamic optimization problems to be well defined (i.e., for solutions to exist), we need to make assumptions about the primitives of the model. Quite generally, we know from basic math—the Weierstrass theorem—that a continuous function attains both a maximum and a minimum when evaluated over a non-empty and compact set. A compact set (of values for a finite vector) means that the set is closed and bounded.[2] Sufficient conditions for this theorem are that (i) $\hat{\mathcal{F}}(y_t)$ is continuous for all y_t, (ii) $h(x_t, y_t)$ is continuous and, in its second argument, strictly monotone for all (x_t, y_t), and (iii) $\Gamma(x_t)$ is non-empty, closed, and bounded for all x_t. Assumption (ii) ensures that we can express y_t as a (continuous) function of (x_t, x_{t+1}), which allows us to write the period objective as a continuous function of this vector. Hence, we have an overall continuous function (of the sequence $\{x_1, x_2, \ldots, x_T, x_{T+1}\}$) to be maximized over a non-empty, compact set.

The two most widely used models in macroeconomics are the consumption-saving model and the neoclassical growth model (NGM). The agent's objective in both models is to maximize lifetime utility choosing the optimal path of consumption. They differ in the production structure and assets available to consumers. In the consumption-saving model, the agent has a time-varying endowment, and can save or borrow at market prices. The NGM, instead, considers the production structure from the Solow model but extends it by endogenizing consumption and investment decisions. In order to fix ideas, it is useful to discuss the main assumptions underlying these models and to map them into our generic formulation.

[1] Often, the problem explicitly says "with x_0 given," to emphasize that it is not a choice variable. However, the fact that it is not a choice variable is already clear since it is not listed among them (under the "max").

[2] By drawing simple graphical examples with non-continuous functions, open sets, and unbounded sets, you can illustrate what can go wrong and why a supremum or infimum may exist but neither a maximum nor a minimum exist.

4.1.1 The consumption-saving model

In the consumption-saving model, there is a representative agent who lives for T periods and must choose the optimal stream of consumption $\{c_t\}_{t=0}^T$, where c_t denotes consumption at time t. We can think of consumption at t as a different good from consumption at $t+1$. Preferences are represented by a utility function $U\left(\{c_t\}_{t=0}^T\right)$. A standard assumption is that this function exhibits "additive separability":

$$U\left(\{c_t\}_{t=0}^T\right) = \sum_{t=0}^{T} \beta^t u\left(c_t\right).$$

Additive separability implies that the marginal utility of consumption at t does not depend on consumption at other times. Notice that the per-period (or instantaneous) utility index $u(\cdot)$ does not depend on time either. The stationary discounting weights satisfy $0 < \beta < 1$, which is consistent with the observation that individuals seem to deem consumption at an early time more valuable than consumption further off in the future. Formally, if consumption were constant over time, i.e., $c_t = c$ for all t, the marginal utility of c_t decreases in t because $u'(c)$ is multiplied by β^t. Of course, consumption in the future can be more valued on the margin if there is less of it, i.e., if it is sufficiently low relative to consumption today.

We assume that the instantaneous utility function $u(c)$ satisfies the following properties.

1. $u(c)$ is strictly increasing.
2. $u(c)$ is strictly concave.
3. $\lim_{c \to 0} u'(c) = \infty$.

The first property states that individuals have positive marginal utility, $u'(c) > 0$. We also assume that marginal utility of consumption is diminishing, or $u''(c) < 0$. The last property is an Inada condition, stating that agents have infinite marginal utility of consumption as c approaches zero.

We abstract from preferences over leisure and, hence, the determination of the number of hours worked (e.g., labor). These will be studied briefly in the next chapter, and more in depth in Chapter 11. For now, we simply assume that individuals have an endowment of one unit of time, and supply it inelastically to production. They are paid a wage w_t, which may vary over time. In addition, they have access to borrowing and lending. We denote their level of *assets* with a_t, with the understanding that $a_t < 0$ indicates that the agent has *debt*. For example, $a_t = 10$ means that another individual owes him or her 10 units of the consumption good, so the agent is a lender, whereas if $a_t = -10$ then the agent owes someone else that amount, so he or she is a borrower. The initial amount of assets is given by a_0. The net interest rate earned on savings is r_t. The budget constraint of the agent can be written as

$$c_t = w_t + (1 + r_t)a_t - a_{t+1}.$$

Borrowing allows the agent to consume more today (in the amount $-a_{t+1}$), but lowers consumption in the future because he or she would need to pay the principal and interest on

the loan, $(1+r_t)a_t$. A standard assumption is that agents cannot borrow more than a certain amount, which is incorporated as the borrowing limit $a_{t+1} \geq \underline{a}$, with $\underline{a}$ being an exogenous constant. In addition, we impose that $a_{T+1} \geq 0$ in order to eliminate the possibility that the agent ends with a positive level of debt in the last period. Absent this constraint, the agent would choose to borrow up to the limit and never pay back, which clearly could not occur in equilibrium (since no lender would be willing to lend the funds in the last period). We will further discuss the issue of the borrowing limit and the terminal condition later in the chapter.

The dynamic optimization problem of the agent can now be written as

$$\max_{\{c_t, a_{t+1}\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u\left(c_t\right) \tag{P2}$$

subject to

$$c_t = w_t + (1 + r_t)a_t - a_{t+1} \quad \forall t \in \{0, ..., T\},$$
$$c_t \geq 0 \quad \forall t \in \{0, ..., T\},$$
$$a_{t+1} \geq \underline{a} \quad \forall t \in \{0, ..., T\},$$

and

$$a_{T+1} \geq 0.$$

The key sequences to be optimally determined are consumption and asset holdings $\{c_t, a_{t+1}\}_{t=0}^{T}$, to maximize the lifetime utility of the agent. We assume that agents know the sequences of prices (wages and interest rates) when making decisions, and take them as given. We defer the discussion of how these are determined in equilibrium to Chapter 5. The main trade-off faced by the agent, then, is whether to consume today or to save (borrow) a unit of consumption in exchange for $1 + r_{t+1}$ additional (less) units in the future.

This model can be mapped to the generic specification as follows: the per-period objective $\hat{\mathcal{F}}\left(\cdot\right)$ corresponds to the instantaneous utility function $u\left(\cdot\right)$, the control variables are given by the consumption sequences $\{c_t\}_{t=0}^{T}$, and the states correspond to the level of assets $x_t = a_t$. The equation relating controls to states is the resource constraint, $h(a_t, c_t) = w_t + (1+r_t)a_t - c_t$. The feasible set is given by $\Gamma(a_t) = [\underline{a}, w_t + (1 + r_t)a_t]$. The lower bound ensures that $a_{t+1} \geq \underline{a}$, whereas the upper bound ensures that savings do not exceed current income so that $c_t \geq 0$ for all t.

4.1.2 The neoclassical growth model

The utility function in the NGM is the same as in the consumption-saving model: an agent wants to maximize his or her discounted lifetime welfare. The assumptions underlying the instantaneous utility function u are the same as in the previous section: the marginal utility of consumption is positive but diminishing, and the Inada condition must hold.

What changes relative to the consumption-saving model is that this is not an endowment economy, but a production economy instead. Rather than focusing on assets delivering an exogenous return, we consider capital accumulation, where returns are determined by the productivity of capital. The production structure is identical to the one in the Solow model.

Following what we learned in Chapter 3, we assume that technology can be represented by the production function $y_t = f(k)$, where we have used the fact that aggregate labor $L = 1$. From the properties of $F(K, L)$, it is easy to show that f is a strictly increasing and strictly concave function in k. In some applications, we relax the assumption to weak concavity in order to accommodate linear production functions. Capital evolves as in the Solow model, i.e.,

$$k_{t+1} = (1 - \delta)k_t + i_t \tag{4.1}$$

with k_0 given. We abstract from population growth and technological progress to simplify the exposition. The resource constraint is $c_t + i_t \leq y_t$, assuming that this is a closed economy where savings are equal to investment. We can combine it with equation (4.1), substituting away investment, to write the resource constraint in terms of consumption and capital:

$$c_t + k_{t+1} \leq f(k_t) + (1 - \delta)k_t. \tag{4.2}$$

There are alternative ways to specify how the markets for labor, consumption, and capital are organized. For example, we could consider firms who own the capital stock and accumulate it over time. They produce output and sell it to consumers, who work in firms. Alternatively, we could assume that agents own capital, make investment decisions, and rent it to firms every period in exchange for a rental rate. These decentralized production structures will be discussed at length in Chapter 5, where we also explain how prices are determined. In Chapter 6, we show that as long as markets are perfectly competitive, the allocations (i.e., quantities of consumption, capital, and investment) in a competitive equilibrium are Pareto optimal and solve the following *planning problem*.

$$\max_{\{c_t, k_{t+1}\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u\left(c_t\right) \tag{4.3}$$

subject to

$$c_t + k_{t+1} \leq f\left(k_t\right) + (1 - \delta)\, k_t \quad \forall t \in \{0, ..., T\}$$

and

$$c_t, k_{t+1} \geq 0 \quad \forall t \in \{0, ..., T\}.$$

The planner chooses allocations directly (note that there are no prices in the equations above) to maximize the lifetime welfare of all agents in the economy. Because all agents are identical, this objective is equivalent to maximizing the welfare of the representative agent. The key sequences to be optimally determined are consumption and capital $\{c_t, k_{t+1}\}_{t=0}^{T}$. The main trade-off faced by consumers is whether to consume today or in the future. Any amount of production that is not consumed can be saved, and therefore invested in new capital, yielding future consumption through the additional amount of goods produced, $f'\left(k_t\right) + 1 - \delta$. The rationale is analogous to that in the consumption-saving model. Because we assume strictly positive marginal utility of consumption, the resource constraint always holds with equality.

We can map this model to the generic specification as follows: the per-period objective $\hat{\mathcal{F}}\left(\cdot\right)$ corresponds to the instantaneous utility function $u\left(\cdot\right)$, the control variables are given by the consumption sequences $\{c_t\}_{t=0}^{T}$, and capital represents the states $x_t = k_t$. The equation relating controls to states is the resource constraint, $h(k_t, c_t) = f\left(k_t\right) + (1 - \delta)\, k_t - c_t$.

Finally, the feasible set is given by $\Gamma(k_t) = [0, f(k_t) + (1-\delta)k_t]$. The lower bound ensures that $k_{t+1} \geq 0$, whereas the upper bound ensures that $c_t \geq 0$. The next step is to discuss how to solve for the optimal sequences.

4.2 Sequential methods: finite horizon

When the horizon is finite, $T < \infty$, it is possible to use the Kuhn-Tucker Theorem to solve a sequential maximization problem (4.3). The associated Kuhn-Tucker conditions are necessary and sufficient for an optimum if the objective function is strictly concave in the choice vector and the constraint set is closed, bounded, and convex. Let's start with the simplest case, in a two-period economy.

4.2.1 A two-period consumption-saving model

Consider the consumption-saving model when $T = 2$, that the agent is borrowing constrained $\underline{a} = 0$, and that initial assets holdings are zero, $a_0 = 0$. Moreover, let us consider a specific example where, for illustration, the endowment profile is decreasing over time, $w_0 \geq w_1$ and the interest rate is time invariant, $r_t = r$. The budget constraints can be written as

$$c_0 = w_0 - a_1 \quad \text{and} \quad c_1 = w_1 + (1+r)a_1,$$

where we already used the fact that the lower bound on assets holdings binds in the last period, $a_2 = 0$. Intuitively, the representative agent will never choose to save in period 1 if the economy ends in period 1. These constraints can be combined by replacing a_1 from the second period constraint into the first period constraint, and re-organizing terms, as follows

$$c_0 + \frac{c_1}{1+r} = w_0 + \frac{w_1}{1+r}.$$

This equation is known as the lifetime budget constraint. It states that the discounted value of lifetime consumption (left-hand side) must be equal to the lifetime value of income (right-hand side). Future periods are discounted by $1+r$, the relative price of consumption between periods 0 and 1. In addition, here we impose the constraint that $a_1 \geq 0$, which can alternatively be written as $w_0 - c_0 \geq 0$. The Lagrangian can be written as

$$\begin{aligned} \mathcal{L} &= u(c_0) + \beta u(c_1) + \mu \left\{ w_0 + \frac{w_1}{1+r} - c_0 - \frac{c_1}{1+r} \right\} \\ &+ \lambda\{w_0 - c_0\}, \end{aligned}$$

where we introduced the Lagrange multiplier μ on the lifetime budget constraint and λ on the non-negativity constraint. We do not consider the non-negativity constraints on consumption here, because the Inada condition $\lim_{c_t \to 0} u'(c_t) = \infty$ ensures that the agent always chooses $c_t > 0$. The first-order conditions with respect to c_0 and c_1 are

$$\frac{\partial \mathcal{L}}{\partial c_0} \quad : \quad u'(c_0) - \mu - \lambda = 0 \tag{4.4}$$

and

$$\frac{\partial \mathcal{L}}{\partial c_1} \quad : \quad \beta u'(c_1) - \mu \frac{1}{1+r} = 0. \tag{4.5}$$

The additional Kuhn-Tucker conditions are

$$\lambda[w_0 - c_0] = 0, \quad w_0 - c_0 \geq 0, \quad \text{and} \quad \lambda \geq 0. \tag{4.6}$$

The first condition in (4.6) is usually referred to as the **complementary slackness condition**. Let us consider an interior solution with $w_0 - c_0 > 0$, implying $\lambda = 0$ from (4.6). It turns out that a decreasing wage path $w_0 > w_1$ and strictly decreasing marginal utility of consumption can guarantee the solution to be interior when $\beta(1+r) \geq 1$. Combining the first-order conditions (4.4) and (4.5) with $\lambda = 0$, we obtain the first-order condition known as the **Euler equation**

$$u'(c_0) = \beta(1+r)u'(c_1).$$

The left-hand side captures the marginal cost of saving an additional unit (which causes a decrease in consumption in the initial period). The right-hand side captures the marginal benefit of saving, which is given by the discounted value of consumption in period $t = 1$ obtained through the returns to savings, $1 + r$. Note that when $\beta(1+r) = 1$, the agent chooses to consume a constant proportion of lifetime earnings every period

$$c_0 = c_1 = \frac{w_0 + \beta w_1}{1+\beta}.$$

Using the budget constraints, this in turn implies that $w_0 - c_0 = a_1 = (w_0 - w_1)/(2+r) > 0$, that is, the solution is indeed interior.[3]

4.2.2 Generic T-period model

When the budget constraint is non-linear in the state variables, it is not possible to construct the lifetime budget constraint as we did in the example above. However, it is still possible to use the Kuhn-Tucker theorem to solve the finite-horizon model, as the next result shows.

Result 1 *Consider a finite-horizon problem (P1), with $\Gamma(x_t) = [\underline{x}, \gamma(x_t)]$. Now use the constraint $x_{t+1} = h(x_t, y_t)$ to solve for y_t as $y_t = \hat{h}(x_t, x_{t+1})$, replace it in the instantaneous objective function, with $\mathcal{F}(x_t, x_{t+1}) \equiv \hat{\mathcal{F}}\left(\hat{h}(x_t, x_{t+1})\right)$, so that we obtain*

$$\max_{\{x_{t+1} \in \Gamma(x_t)\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t \mathcal{F}(x_t, x_{t+1}).$$

Suppose that $\mathcal{F}(x_t, x_{t+1})$ is increasing in its first argument, decreasing in its second argument, continuously differentiable, and jointly concave in (x_t, x_{t+1}). If

[3] In the general case, we can proceed by first finding an interior *candidate* solution and then verifying whether or not a_1 satisfies its constraint (in this case that a_1 be non-negative); if this works, it is a solution—given the strictly concave utility function and convex constraint set. If a_1 violates its constraint, set it equal to its boundary value and solve for λ, in order to verify that it is non-negative.

(i) $\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta\mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) = 0,\ \forall t < T$ *with* $\left\{x_{t+1}^*\right\}_{t=0}^{T-1} \in int\Gamma(x_t)$*, where* $\mathcal{F}_i(x_t, x_{t+1})$ *represents the partial derivative with respect to* i*th argument.*

(ii) $x_{T+1}^* = \underline{x}$*,*

then the sequence $\left\{x_{t+1}^*\right\}_{t=0}^{T}$ *maximizes the objective.*

We omit the proof of this result here. In our section on infinite-horizon optimization below, we state the corresponding result for that setting and, in our appendix, prove it from first principles. That proof can easily be adapted to the present, finite-horizon case.

The result states that it is optimal to choose the lower bound of the feasible set in the last period, $\underline{x}$, and that an interior solution for $t < T$ satisfies the first-order condition (i), which represents the Euler equation for the general model. This equation is sometimes referred to as a "variational" condition (as part of "calculus of variation"): given two boundary conditions x_t and x_{t+2}, we vary the intermediate value x_{t+1} so as to achieve the best outcome. Combining these variational conditions, we notice that there is a total of $T+2$ equations and $T+2$ unknowns, namely the sequence of states, plus the initial and terminal conditions. This is called a *difference equation* in the sequence of state variables. It is a *second-order* difference equation because there are two lags of x in the equation. Since the number of unknowns is equal to the number of equations, the difference equation system will typically have a solution.

4.2.3 The finite-horizon neoclassical growth model

In the NGM, we have that $\underline{x} = 0$. Solving for consumption from the resource constraint and inserting it into the objective, the problem can be written more compactly as

$$\max_{\{k_{t+1}\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u\left(f(k_t) + (1-\delta)\,k_t - k_{t+1}\right) \tag{P3}$$

subject to

$$k_{t+1} \geq 0 \quad \forall t \leq T.$$

Our assumptions on the utility and production functions ensure concavity of the objective function and a closed, bounded, and convex constraint set, so that the assumptions guaranteeing that there is a unique maximizer are fulfilled.[4] As in the two-period consumption-saving model, we are omitting the constraint $c_t \geq 0$, given the Inada condition in the utility function.

The Lagrangian associated to our maximization problem (P3) is

$$\mathcal{L} = \sum_{t=0}^{T} \beta^t \left\{u(f(k_t) + (1-\delta)\,k_t - k_{t+1}) + \mu_t k_{t+1}\right\},$$

[4]To show that the period objective is concave in (k_t, k_{t+1}), it is necessary to go beyond concavity in k_t and k_{t+1} separately: one needs to check that the Hessian is globally negative-definite.

where we introduced the Lagrange/Kuhn-Tucker multipliers $\beta^t \mu_t$ corresponding to the non-negativity constraints on capital for each period t. The first-order conditions are

$$\frac{\partial \mathcal{L}}{\partial k_{t+1}} \quad : \quad -u'(c_t) + \beta u'(c_{t+1})[f'(k_{t+1}) + 1 - \delta] + \mu_t = 0, \quad t \in \{0, ..., T-1\} \tag{4.7}$$

and

$$\frac{\partial \mathcal{L}}{\partial k_{T+1}} \quad : \quad -u'(c_T) + \mu_T = 0. \tag{4.8}$$

The first-order condition in the final period is different from the ones in earlier periods because the economy ends at that time. Finally, the Kuhn-Tucker conditions also include

$$\mu_t k_{t+1} = 0, \ t \in \{0, ..., T\}, \tag{4.9}$$
$$k_{t+1} \geq 0, \ t \in \{0, ..., T\},$$

and

$$\mu_t \geq 0, \ t \in \{0, ..., T\}.$$

Equation (4.9) parallels the first condition of (4.6) in the two-period consumption-savings problem (complementary slackness condition). Because we assume that $u(c_t)$ is strictly increasing, equation (4.8) implies that $\mu_T > 0$. From (4.9) evaluated at $t = T$, we find that $k_{T+1} = 0$. This result establishes the terminal condition for our maximization problem: consumers leave no capital for production after the final period, since they receive no utility from that capital and would rather use it for consumption during their lifetime. The insight is trivial here, but will be relevant also when we consider an infinite-horizon economy.

Given the Inada conditions, in particular $\lim_{k \to 0} f'(k) = \infty$ and $\lim_{c \to 0} u'(c) = \infty$, it is optimal to set $k_{t+1} > 0$ for all $t < T$. Hence, the non-negativity constraint on capital will never be binding, implying that $\mu_t = 0$ for $t < T$. Replacing this result in equation (4.7) delivers the Euler equation,

$$u'\left[f\left(k_t\right) + (1-\delta)k_t - k_{t+1}\right] = \beta u'\left[f\left(k_{t+1}\right) + (1-\delta)k_{t+1} - k_{t+2}\right]\left[f'\left(k_{t+1}\right) + 1 - \delta\right], \tag{4.10}$$

which holds for all $t \in \{0, ..., T-1\}$. This equation, together with the initial condition k_0 and the terminal condition derived before, $k_{T+1} = 0$, determines the capital sequence $\{k_{t+1}\}_{t=0}^{T}$ (e.g., $T+2$ equations and $T+2$ unknowns). Because the first-order conditions are sufficient in the example, there is a unique solution to the difference equation (4.10) describing the evolution of capital over time.

To interpret the key equation for optimization, it is useful to break the Euler equation down into three components:

$$\underbrace{u'\left(c_t\right)}_{\substack{\text{marginal cost} \\ \text{of investment}}} = \underbrace{\beta u'\left(c_{t+1}\right)}_{\substack{\text{utility increase} \\ \text{per unit return}}} \cdot \underbrace{\left[f'\left(k_{t+1}\right) + 1 - \delta\right]}_{\text{return on investment}}.$$

The left-hand side represents the marginal cost of investing one unit of consumption today, generating a disutility loss of $u'(c_t)$. The right-hand side represents the marginal benefit of this investment. Higher investment increases capital next period, which produces additional output (plus un-depreciated capital) $f'(k_{t+1}) + 1 - \delta$. This return increases consumption next period, with a discounted per-unit utility gain of $\beta u'(c_{t+1})$.

4.2.4 Solving a finite-horizon model

We have seen how to derive the set of equations that will determine the solution to the maximization problem. How does one solve these equations, however? Several methods are available. The first one uses "backward induction": starting at the final period, T, we can use the resource constraint and the Euler equation iteratively moving backwards. This method is illustrated in the following example, which is designed to yield closed-form solutions at each stage (it is the only known example with strictly concave utility and production functions that can be solved analytically).

Example 4.1 *Consider a T-period economy where utility is logarithmic, $u(c) = \log c$, the production function is Cobb-Douglas $f(k) = Ak^{\alpha}$, and there is full depreciation $\delta = 1$. The Euler equation (4.10) for $t < T$ becomes*

$$\frac{1}{Ak_t^{\alpha} - k_{t+1}} = \beta \frac{1}{Ak_{t+1}^{\alpha} - k_{t+2}} \alpha A k_{t+1}^{\alpha-1}. \tag{4.11}$$

The last term is the marginal product of capital $f'(k) = \alpha A k^{\alpha-1}$. Evaluating equation (4.11) at period $t = T - 1$, replacing the terminal condition $k_{T+1} = 0$ in its right-hand side, and simplifying delivers

$$k_T = \frac{\alpha\beta}{1 + \alpha\beta} A k_{T-1}^{\alpha}.$$

We can now use this result, together with equation (4.11) evaluated at period $T-2$, to obtain

$$k_{T-1} = \frac{\alpha\beta(1 + \alpha\beta)}{1 + \alpha\beta + (\alpha\beta)^2} A k_{T-2}^{\alpha}.$$

Going backwards in time, it is possible to see that

$$k_{T-t} = \frac{\alpha\beta(1 + \alpha\beta + ... + (\alpha\beta)^t)}{1 + \alpha\beta + ... + (\alpha\beta)^{t+1}} A k_{T-t-1}^{\alpha}.$$

Using the fact that $\alpha\beta < 1$, we can use the properties of geometric series in the numerator and the denominator to simplify this expression. That, together with a change of variables, allows us to obtain a formula that describes the evolution of capital in closed form:

$$k_{t+1} = \alpha\beta \frac{1 - (\alpha\beta)^{T-t}}{1 - (\alpha\beta)^{T-t+1}} A k_t^{\alpha}.$$

Consumption becomes

$$c_t = \frac{1 - \alpha\beta}{1 - (\alpha\beta)^{T-t+1}} A k_t^{\alpha}.$$

There are a few characteristics of the solution that we would like to highlight. First, we obtained an expression that, given the initial condition k_0, fully describes the evolution of capital, output, and consumption for the whole time horizon: the outcomes depend explicitly on parameters and on k_0. Second, we see that it is optimal to save (and invest) a proportion

$$s_t = \alpha\beta \frac{1 - (\alpha\beta)^{T-t}}{1 - (\alpha\beta)^{T-t+1}}$$

of output every period. In contrast to the assumption of the Solow model, the optimal saving rate depends on time (the time left to the final date T). But, like in the Solow model, the saving rate does not depend on the level of the capital stock. Two key parameters determine the saving rate: the discount factor and the degree of concavity of the production function. Third, although the utility function is strictly concave, consumption is not fully smoothed: it will, in general, vary over time. The reason for the lack of full smoothing is that the level of the capital stock influences the marginal return on saving when the production function is neoclassical: the higher the capital stock, the lower is this return. The Euler equation tells us that the higher the marginal return on saving, the higher the consumption growth should be, implying that for capital stocks below (above) steady state, consumption rises (falls) over time.[5]

It is possible to obtain analytical solutions in a few other cases; one is where the production function is linear and preferences are represented by a u as given in (2.5); this function is often called the Constant Relative Risk Aversion (CRRA) utility function.

Constant Relative Risk Aversion (CRRA) utility function

The CRRA function is one of the most commonly used additively separable utility functions in macroeconomics[a]

$$u(c) = \frac{c^{1-\sigma} - 1}{1-\sigma} \quad \text{where } \sigma \geq 0 \text{ and } \sigma \neq 1$$

It has, as special cases,

$$\begin{array}{ll} \sigma = 0 & \text{linear utility,} \\ \sigma > 0 & \text{strictly concave utility,} \\ \sigma \to 1 & \text{logarithmic utility.} \end{array}$$

The limit case where $\sigma \to \infty$ is the zero function. However, the relevant limit should be seen as that obtained by raising $\sum_{t=0}^{\infty} \beta^t c_t^{1-\sigma}$ to a power $1/(1-\sigma)$, in which case the limit becomes a "Leontief function": $\min_t \{c_t\}_{t=0}^{\infty}$.[b]

[a] We do not consider uncertainty in this chapter, so the concept of the "relative risk aversion" is not directly relevant. Here this term serves only as a label for a class of utility functions.

[b] The stated utility function here is a monotone transformation of the original function and we know that the behavior given by a utility function is preserved under monotone transformations.

We define the elasticity of intertemporal substitution (EIS) as the percentage change in consumption between periods t and $t+s$ in response to a percentage change in the returns to investment between the same two periods, assuming optimizing behavior subject to a standard budget constraint, taking other returns and the agent's income as given.[6] Thus,

[5] The special case $\alpha = 1$ means that the marginal return on saving is always A. Then, unless $\beta A = 1$, the economy will grow or shrink over time at a constant rate, as will consumption. Consumption, along with capital, will thus only be constant if $\beta A = 1$.

[6] That is, we use the notion of an *uncompensated* elasticity. Also, our notion presumes the absence of other goods in the utility function, in which case one needs to specify how the choices of the other goods are

we have

$$EIS \equiv \frac{\frac{\partial(c_{t+s}/c_t)}{c_{t+s}/c_t}}{\frac{\partial R_{t,t+s}}{R_{t,t+s}}}.$$

In the next example, we show that the CRRA function has a constant intertemporal elasticity of substitution, which is equal to $1/\sigma$. For this reason, a CRRA utility function is also sometimes referred to as a Constant Elasticity of Intertemporal Substitution (CEIS) utility function.

Example 4.2 *Consider a T-period economy where utility is CRRA, $u(c) = \frac{c^{1-\sigma}-1}{1-\sigma}$ and the production function is linear $f_t(k_t) = R_t k_t$. In this case, the Euler equation reads*

$$u'(c_t) = \beta u'(c_{t+1}) R_{t+1}.$$

Repeated substitution delivers

$$\begin{aligned} u'(c_t) &= \beta^s u'(c_{t+s}) \underbrace{R_{t+1}R_{t+2}...R_{t+s}}_{\equiv R_{t,t+s}} \\ u'(c) &= c^{-\sigma} \Rightarrow c_t^{-\sigma} = \beta^s c_{t+s}^{-\sigma} R_{t,t+s} \\ \frac{c_{t+s}}{c_t} &= (\beta^s)^{\frac{1}{\sigma}} (R_{t,t+s})^{\frac{1}{\sigma}}. \end{aligned}$$

This means that the EIS becomes

$$\frac{\frac{\partial(c_{t+s}/c_t)}{c_{t+s}/c_t}}{\frac{\partial R_{t,t+s}}{R_{t,t+s}}} = \frac{\partial \log(c_{t+s}/c_t)}{\partial \log R_{t,t+s}} = \frac{1}{\sigma}.$$

When $\sigma \to 1$, the relative expenditure shares $c_t/(c_{t+s}/R_{t,t+s})$ do not change: this corresponds to the logarithmic case. When $\sigma > 1$, an increase in $R_{t,t+s}$ would lead c_t to increase and investment to decrease: the income effect, leading to smoothing across all goods, is larger than the substitution effect. Finally, when $\sigma < 1$, the substitution effect is stronger than the income effect: investment rises whenever $R_{t,t+s}$ increases. When $\sigma = 0$, the elasticity is infinite and investment responds discontinuously to $R_{t,t+s}$.

Another, and absolutely central, reason that the CRRA utility function plays an important role for us is that it is the only one that is consistent with balanced growth, as described in Section 4.3.3.

4.3 Sequential methods: infinite horizon

In this section, we extend our model to infinite periods. The main advantage of an infinite horizon is that the household problem becomes stationary: the maximization problem at date t is exactly the same as in period $t+1$ (for a given starting level of capital). This

made as the returns are changed.

property is in contrast to that in the previous section, where decisions were significantly affected by how many periods the individual had left (see Example 4.1). A large number of macroeconomic applications, particularly those studying the long-run evolution of aggregate economic variables, use infinite-lived agents as their main building block.

For the typical models that macroeconomists use, the infinite-horizon version behaves very similarly to the finite-horizon version when the latter's remaining time horizon is long enough.[7] However, let us also discuss whether an infinite time horizon is a sensible assumption: after all, people do not live forever. However, to the extent that individuals are altruistic, they care about their descendants. Let $u(c_t)$ denote the utility flow to generation t. We can then interpret β^t as the weight individuals attach to the utility enjoyed by their descendants, t generations down the family tree. Their total welfare is given by $\sum_{t=0}^{\infty} \beta^t u(c_t)$. As long as $\beta < 1$, agents care more about themselves than about their offspring.[8]

4.3.1 Mathematical considerations

Because agents are now choosing infinite sequences of consumption and investment, models with an infinite time horizon demand more advanced mathematical tools.

A basic question is whether the solution to the planner's problem exists once choices are no longer elements of the Euclidean space $\Re^T$. In more general notation, suppose we are seeking to maximize a function $U(x)$, where $x \in S$ and S is a set that includes infinite sequences. If U is continuous, we can invoke the Weierstrass theorem, provided that the set S is nonempty and compact. For finite sequences, continuity and compactness are defined in standard ways. But for infinite-dimensional sequences, several issues arise. How do we define continuity in this setup? What is an open set? What does compactness mean? Answering these questions in detail is beyond the scope of this book; we refer you, for example, to Stokey and Lucas (1989). For illustration, we will, however, provide some specific examples where the maximization problem may be ill-defined (i.e., have no solution) unless a set of necessary conditions holds.

Unbounded utility Continuity of the objective requires boundedness. A necessary condition for the lifetime utility U to be bounded is that consumption streams do not yield "infinite" utility. If two consumption streams do so, they cannot be compared and the maximization problem is ill-defined. For example, consider a plan specifying equal amounts of consumption each period, $\{c_t\}_{t=0}^{\infty} = \{\bar{c}\}_{t=0}^{\infty}$, delivering $U = \sum_{t=0}^{\infty} \beta^t u(\bar{c})$. Clearly, this function is unbounded (e.g., does not have a finite limit) unless $\beta < 1$.

This may not be sufficient, however. Suppose that the constraints allow for a constantly increasing consumption stream $\{c_t\}_{t=0}^{\infty} = \left\{c_0 (1+\gamma)^t\right\}_{t=0}^{\infty}$. The lifetime utility is now $U = \sum_{t=0}^{\infty} \beta^t u\left(c_0 (1+\gamma)^t\right)$. Even if $\beta < 1$ (so β^t is decreasing to 0), the argument inside the utility function is growing at rate γ. The shape of the utility function, hence, is key to

[7] Game theory is a sharp contrast here: we know that an infinite horizon can then open up to the existence of many equilibria, e.g., the trigger-strategy outcomes in repeated games.

[8] In this simple example we attach only one consumption level to each generation. The example could be extended to the case where people of generation t consume in multiple periods; though more complicated to formulate, this extension would be straightforward and the main insights would carry over.

determining whether the maximization problem is well defined. In the case of a CRRA utility function $u(c) = (c^{1-\sigma} - 1)/(1-\sigma)$, we obtain a level of utility that is a geometric sum. Hence, boundedness requires $\beta(1+\gamma)^{1-\sigma} < 1$. If $\sigma < 1$, so that there is less than logarithmic curvature, this requirement involves an upper bound on γ for utility to be a positive, finite number. If $\sigma > 1$, positive growth ($\gamma > 0$) implies that the boundedness condition is met; with negative growth at a sufficiently high rate, it will not be.[9]

Constraint sets that are "too large" The problem can be ill-defined if the constraint sets are "too large". In the consumption-saving model discussed at the outset of the chapter, we imposed the constraint that borrowing could not exceed the exogenous amount $\underline{a}$. This condition ensures that the constraint set is bounded, and hence that the problem is well defined.

In some applications, however, imposing the constraint $a_{t+1} \geq \underline{a}$ for each t can be "too restrictive"; it will rule out feasible borrowing. Instead, for example, we could only impose that $a_{T+1} \geq 0$, i.e., that the agent cannot borrow in the very last period. This terminal constraint is important: if this constraint was not imposed, any individual would have incentives to accumulate an infinite amount of debt going into the final period. Such a restriction also makes sense: knowing that the consumer would engage in such a scheme, no lender would be willing to lend at any positive rate, as the loan would be defaulted with probability 1.

With an infinite horizon, a final-period constraint loses meaning in a literal sense. Instead, we impose a constraint that is its appropriate infinite-period extension, known as the **no Ponzi game** (nPg) condition. In words, this condition rules out "borrowing at infinity, measured in present value." This requirement represents a restriction on the agent's constraint set, preventing it from being so large as to allow the agent to attain arbitrarily high utility. To see how the condition comes about, let us look at a simple example.

Suppose we endow a consumer with a given initial amount of net assets, a_0, representing claims against other agents. Additionally, suppose that the agent has no other sources of income, so the budget constraint is

$$c_t + a_{t+1} = (1+r)a_t, \forall t \geq 0,$$

where we assume that $r > 0$.[10] Consider a candidate solution to consumer's maximization problem $\{c_t^*\}_{t=0}^{\infty}$. Absent further constraints, the agent could improve on $\{c_t^*\}_{t=0}^{\infty}$ as follows:

1. Let $\tilde{c}_0 = c_0^* + \epsilon$, with $\epsilon > 0$, thus making $\tilde{a}_1 = a_1^* - \epsilon$.

2. For every $t \geq 1$ leave $\tilde{c}_t = c_t^*$ by setting $\tilde{a}_{t+1} = a_{t+1}^* - \epsilon(1+r)^t$.

Given a strictly increasing utility function, the agent is clearly better off under this alternative consumption allocation, which satisfies the budget constraint period by period. Because this sort of improvement is possible for *any* candidate solution, there cannot be a maximum

[9] Negative growth, i.e., $\gamma < 0$, does not occur in the context of our standard balanced-growth model, but it can hypothetically occur if natural resources are assumed to be finite and essential for production. In this case, utility could hence become unboundedly negative—recall that u is negative in this case—and there may even be no allocation of resources with finite utility. For more, see Chapter 25.

[10] Cases with $r \leq 0$ can sometimes be relevant but we do not consider them here.

for lifetime utility.[11] Note that with this alternative allocation, the agent's debt is growing without bound at rate $1+r$, and it is never repaid. This type of scheme, borrowing $(1+r)a_t$ every period t to keep rolling the debt, is often called a "Ponzi scheme" or a "Ponzi game". It is crucial in the infinite-horizon model to impose a constraint that rules out the Ponzi scheme. A condition that works is the nPg condition in the following form:

$$\lim_{t\to\infty} \frac{a_{t+1}}{(1+r)^t} \geq 0. \tag{4.12}$$

Intuitively, the agent cannot engage in borrowing and lending so that their "terminal asset holdings" (in present-value terms) are negative, because this means that they would borrow and not pay back.

We can use the nPg condition to simplify, or *consolidate*, the sequence of budget constraints. To do this, solve for a_1 from the budget constraint in period 1 and substitute the expression into the budget constraint in period 0: we obtain a budget containing c_0, c_1, a_0, and a_2. Next, replace a_2 in this expression by solving for it in the next budget constraint, and proceed this way forward. After T substitutions, we obtain

$$\sum_{t=0}^{T} c_t \frac{1}{(1+r)^t} = a_0(1+r) - \frac{a_{T+1}}{(1+r)^T}.$$

Taking limits, we arrive at

$$\sum_{t=0}^{\infty} c_t \frac{1}{(1+r)^t} = a_0(1+r) - \lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} \leq a_0(1+r), \tag{4.13}$$

where the inequality comes from the nPg condition. This is the lifetime budget constraint in the infinite-horizon model.

We motivated ruling out Ponzi schemes as a natural infinite-horizon extension of the constraint $a_{T+1} \geq 0$ in a finite-horizon model. There is also a constraint often labeled the "natural borrowing limit." It can be used both in the finite-horizon and the infinite-horizon cases and it captures, precisely, the notion of a loosest possible constraint on borrowing: the only restriction is that you are able to pay back, in a present-value sense, if you set consumption to zero at all future times. In the particular model here, the natural borrowing limit is zero at all times, but in general the constraint depends on the future stream of non-asset income. This case is analyzed in Appendix 4.A.1, which shows that imposing the natural borrowing limit, imposing the lifetime budget constraint, and imposing the nPg condition are all equivalent.

Note that the lifetime budget constraint is often written with equality, i.e.,

$$\sum_{t=0}^{\infty} c_t \frac{1}{(1+r)^t} = a_0(1+r), \tag{4.14}$$

because, as we shall see below, $\lim_{T\to\infty} a_{T+1}/(1+r)^T$ will never be chosen to be positive, so long as the utility function is strictly increasing. The reason is that assets do not themselves

[11] One could imagine a maximum if an arbitrary upper bound is placed on consumption at each date, but such ad-hoc assumptions are undesirable.

contribute to utility, so it is always better to increase consumption as long as it is positive. The inequality in the opposite direction of (4.12),

$$\lim_{T\to\infty}\frac{a_{T+1}}{(1+r)^T}\leq 0, \tag{4.15}$$

therefore, is a part of the optimality condition. The condition (4.15) will be referred to as the **transversality condition** (TVC). The equality (4.14) is the combination of the inequalities (4.13) (which comes from the nPg condition (4.12)) and the TVC (4.15). The nPg and the TVC are different in nature: the former is a restriction on what paths the consumer is allowed to choose whereas the latter is a self-imposed condition—it is chosen by the consumer.

Let us now work out a full solution to the problem above.

Example 4.3 *Consider the infinite-horizon version of the consumption-saving model (without borrowing constraints) given a logarithmic utility function* $u(c)=\log c$. *The optimization problem is:*

$$\max_{\{c_t,a_{t+1}\}_{t=0}^{\infty}}\sum_{t=0}^{\infty}\beta^t\log c_t$$

subject to

$$c_t+a_{t+1}=a_t(1+r)\quad\forall t\geq 0$$

and the nPg condition.

To solve this problem, replace the period budget constraints with a consolidated one evaluated at equality (hence, already using the TVC, i.e., the fact that it is not optimal to save a positive amount, in present-value terms, at infinity):

$$\sum_{t=0}^{\infty}c_t\left(\frac{1}{1+r}\right)^t=a_0(1+r).$$

With this simplification, the first-order conditions are

$$\beta^t\frac{1}{c_t}=\lambda\left(\frac{1}{1+r}\right)^t,\forall t\geq 0,$$

where λ *is the Lagrange multiplier associated with the consolidated budget constraint. From the first-order conditions it follows that*

$$c_t=[\beta(1+r)]^t c_0,\forall t\geq 1.$$

Substituting this expression into the consolidated budget constraint, we obtain

$$\begin{aligned}\sum_{t=0}^{\infty}\beta^t(1+r)^t\frac{1}{(1+r)^t}c_0 &= a_0(1+r)\\ c_0\sum_{t=0}^{\infty}\beta^t &= a_0(1+r).\end{aligned}$$

From here, $c_0=a_0(1-\beta)(1+r)$, *and consumption in the periods* $t\geq 1$ *can be recovered from* $c_t=[\beta(1+r)]^t c_0$.

Sufficient conditions: the transversality condition (TVC) In general, the infinite-horizon maximization problems involve the same mathematical techniques as the finite-horizon ones. In particular, we make use of (Kuhn-Tucker) first-order conditions. In the neoclassical growth model, these lead to a second-order difference equation, the Euler equation, defining a path for the state variable given the initial condition k_0. But unlike in the finite horizon case, where it was optimal to set $k_{T+1} = 0$, there is no final condition that allows us to pin down the sequence of capital. Therefore, the difference equation that characterizes the first-order condition may have an infinite number of solutions. To determine the solution, we need to make use of an additional optimality condition that we already briefly discussed: the *transversality condition*. This condition, which we will now discuss in more general terms, captures the principle that it cannot be optimal for an agent to choose a sequence of capital involving, in present-value utility terms, a positive shadow value as $t \to \infty$. In the consumption-saving problem such behavior is clearly sub-optimal: lower savings would be feasible and yield higher lifetime utility.

We will not prove the necessity of the TVC here. We will, however, provide a sufficiency condition for a generic optimization problem. We will, moreover, offer a proof strategy that also allows us to derive what form the TVC must take (it is not always obvious what its precise form should be). The message of the following proposition is that if we have a convex maximization problem (utility is concave and the constraint set convex) and a sequence $\{x_{t+1}\}_{t=0}^{\infty}$ that satisfies the Kuhn-Tucker first-order conditions and the transversality condition, then indeed we have a maximum. Formally, we have the following.

Proposition 4.4 *Consider the infinite-horizon version of the maximization problem* (P1)*, where we use* $x_{t+1} = h(x_t, y_t)$ *to replace* y_t *in the objective:*

$$\max_{\{x_{t+1} \in \Gamma(x_t)\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \mathcal{F}(x_t, x_{t+1}).$$

If $x_{t+1}^* \in$ *int* $\Gamma(x_t)$ *for all* t*,*

(i) Euler equation: $\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta \mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) = 0\ \forall t$

(ii) TVC: $\lim_{t\to\infty} \beta^t \mathcal{F}_1\left(x_t^*,\ x_{t+1}^*\right) x_t^* = 0,$

$\mathcal{F}\left(x_t, x_{t+1}\right)$ *is jointly concave in* (x_t, x_{t+1}) *and increasing in its first argument, and* $\Gamma(x)$ *is a convex set for all* x*, then* $\left\{x_{t+1}^*\right\}_{t=0}^{\infty}$ *maximizes the objective.*

Proof. See Appendix 4.A.3 ■

The proof is relatively mechanical: it repeatedly uses the Euler equation, (i), and a definition of concavity, namely, that a concave function is always globally below its tangent hyperplane. It thus shows that the stated sequence gives a higher value to the objective than any other feasible sequence so long as a remaining inequality is met. That remaining inequality is met when the TVC is satisfied: condition (ii) of the theorem.

In the neoclassical growth model, the partial derivative $\mathcal{F}_1\left(x_t, x_{t+1}\right)$ becomes $u'(c_t)[f'(k_t)+1-\delta]$, which corresponds to the marginal utility of increasing capital in period t. The TVC thus reads

$$\lim_{t\to\infty} \beta^t u'(c_t)[f'(k_t) + 1 - \delta]k_t = 0.$$

It states that the discounted present-value of each additional unit of capital times the stock of capital has to be zero in the limit. If this requirement is not met, it would be beneficial to modify the path of capital in order to increase consumption, generating higher lifetime utility, without violating feasibility.

In cases where a proposed chosen sequence becomes stationary in the limit—such as in the neoclassical growth model without growth—the TVC is satisfied "automatically" so long as $\beta < 1$. When consumption and capital grow, i.e., due to technical change, the condition is less trivial. Under the form of utility that allows balanced growth, $u(c) = (c^{1-\sigma} - 1)/(1-\sigma)$, $\beta^t u'(c_t)k_t$ will grow at the gross rate $\beta\gamma^{1-\sigma}$, since both c_t and k_t grow at gross rate γ on a balanced path. The transversality condition thus requires $\beta\gamma^{1-\sigma} < 1$, which we recall is also the condition that makes utility bounded.

It is yet again worth emphasizing that the transversality condition and the no-Ponzi game conditions are conceptually distinct. The TVC (jointly with the Euler equation being met at all times) is a sufficient condition for optimization. It states that the value of assets cannot be positive in the limit, because the agent could otherwise be made better off by reducing them. Hence, it is eliminating sequences that cannot be optimal by ruling out over-accumulation of wealth.[12] The nPg condition, on the other hand, is an institutional constraint ensuring that the agent cannot have positive debt (in present value terms) in the limit. An agent would want to choose to increase debt unboundedly if he or she were allowed to violate it. The latter could not reasonably be thought to occur in a market economy and therefore the constraint is imposed. In sum, the TVC is a self-imposed constraint ruling out over-accumulation of assets, whereas the nPg condition is an externally imposed constraint ruling out of extreme under-accumulation (exploding debt).

4.3.2 Solving the infinite-horizon neoclassical growth model

The infinite-horizon maximization problem can be solved taking the limit of the solution we found in the finite horizon case, and making sure that the transversality condition is met. A question, to which we will return later, is whether this model, like the Solow model, will imply global convergence to a steady state. A preliminary inquiry involves finding the set of steady states. This is straightforward: the Euler equation (4.10) implies that for any constant positive level of consumption—which means that $u'(c_t) = u'(c_{t+1})$ can be eliminated from the equation—we obtain

$$1 = \beta(f'(\bar{k}) + 1 - \delta). \tag{4.16}$$

Given that f is strictly concave and satisfies Inada conditions, this implies a unique solution.

We will return to the general formulation later, once we have covered recursive methods, and then establish global convergence to $\bar{k}$ from any positive starting level of capital. For now, we will look at an example where we can solve for dynamics explicitly.

Example 4.5 *Consider an infinite-horizon economy where utility is logarithmic, $u(c) = \log c$, the production function is Cobb-Douglas $f(k) = Ak^\alpha$, and there is full depreciation: $\delta = 1$. Truncating the economy at $t = T$ delivers the same optimality conditions and analytical*

[12] The nature of this argument is that TVC is a necessary condition for optimization, which it is under some conditions; we do not prove this here.

expression for capital next period that we found in Example 4.1. Denote the solution to the truncated problem with k_{t+1}^T. Taking limits, we can find a candidate solution to the infinite horizon problem k_{t+1} as

$$\begin{aligned} \lim_{T\to\infty} k_{t+1}^T &= \lim_{T\to\infty} \alpha\beta \frac{1-(\alpha\beta)^{T-t}}{1-(\alpha\beta)^{T-t+1}} A k_t^\alpha \\ k_{t+1} &= \alpha\beta A k_t^\alpha \end{aligned} \tag{4.17}$$

To check that the TVC holds, as pointed out above, we only need to look at a limit and not at other aspects of the sequence for capital. So, (i), (4.17) implies that the capital sequence converges to a limit (as will, then, consumption) and, (ii) therefore the limit behavior of the expression $\beta^t \mathcal{F}_1(k_t,\ k_{t+1})\, k_t$ will boil down to a constant times β^t. Thus, the TVC will be met. To see the concrete expressions, first note that $\mathcal{F}_1(k_t,\ k_{t+1}) = u'(c)f'(k_t)$, since consumption is simply $c = f(k_t) - k_{t+1}$. The TVC can be written as

$$\lim_{t\to\infty} \beta^t u'(c_t) f'(k_t) k_t = \lim_{t\to\infty} \beta^t \frac{A\alpha k_t^{\alpha-1}}{A k_t^\alpha - k_{t+1}} k_t.$$

Using equation (4.17) and simplifying delivers $\lim_{t\to\infty} \beta^t \alpha/(1-\alpha\beta) = 0$, since $\beta < 1$.

Another method typically used to solve infinite horizon models analytically is called "guess and verify." The strategy consists of guessing a generic functional form for k_{t+1} as a function of k_t and using the Euler equation to verify that the guess is correct. We illustrate this method in the following example.

Example 4.6 *Consider the economy from Example 4.5. Guess that $k_{t+1} = sAk_t^\alpha$, where s is an unknown parameter (the saving rate). Insert this guess into the Euler equation (4.10) to obtain*

$$\frac{1}{Ak_t^\alpha - sAk_t^\alpha} = \beta\alpha A \frac{k_{t+1}^{\alpha-1}}{Ak_{t+1}^\alpha - sAk_{t+1}^\alpha}.$$

After some manipulations, we can verify that $s = \alpha\beta$ satisfies the equation above no matter what k_t is.

An obvious complication with this approach is of course how to come up with an insightful initial guess. A standard procedure is to use a combination of the truncation-at-T method and the guess and verify method. In other words, it is possible to solve the problem backwards a couple of periods to get a sense of the possible functional form, and then verify whether this guess is correct using the Euler equation.

It is worth noticing that, in the infinite-horizon economy of Example 4.5, the stock of capital evolves as assumed by the Solow model: agents invest a *constant* proportion, $s = \alpha\beta$, of their income every period. Moreover, using what we have learned in Chapter 2, we can show that this economy converges to a steady state $\bar{k}$. To compute it, simply evaluate our solution at the steady state, $k_t = k_{t+1} = \bar{k}$,

$$\bar{k} = \beta\alpha A \bar{k}^\alpha \ \Rightarrow\ \bar{k} = (\beta\alpha A)^{\frac{1}{1-\alpha}},$$

which is identical to what we found in Chapter 3, given $\delta = 1$.

In general, when closed-form solutions are not possible to solve for, we need to resort to numerical methods. It is possible to solve for sequences numerically, but seeking numerical solutions using dynamic programming is also possible. Dynamic programming also delivers conceptual insights. The following section turns to this method.

In the case of a relatively simple model like the NGM, yet another method of characterizing the dynamics is to use *phase diagrams*. The solution to a general NGM can be written as a set of two difference equations for two variables (k_t, c_t):

$$k_{t+1} - k_t = f(k_t) - \delta k_t - c_t$$

and

$$u'(c_t) = \beta u'(c_{t+1})(f'(k_{t+1}) + 1 - \delta).$$

The first is the resource constraint for the social planner, and the second is the Euler equation. We can look for the dynamic path of (k_t, c_t) that satisfies these two difference equations, together with the nonnegativity constraints for both variables and the TVC. The graphical analysis using the phase diagram is explained in Appendix 4.A.4.

4.3.3 Balanced growth in the neoclassical growth model

The economies described in the previous sections eventually converge to a steady state, exhibiting no growth in the long run. It is possible to study a version of the NGM in which there is balanced growth, with a production structure similar to the one studied in Chapter 3.2. There, we considered $Y_t = F(K_t, A_t L_t)$, where $A_t L_t$ denotes "efficiency units of labor." An important assumption that allows us to obtain balanced growth is that L_t and A_t grow at the (constant) rates n and γ, respectively. Because of population growth, we now need to take a stand on the welfare criterion by giving weight to the current population and people born in the future. In what follows, we assume that the social planner is "utilitarian" and maximize the sum of utility in the entire economy:

$$\sum_{t=0}^{\infty} \beta^t L_t u\left(c_t\right),$$

where $c_t = C_t / L_t$ is the per capita consumption. An alternative formulation is to assume the social planner cares about the "per capita" utility in each period: $\sum_{t=0}^{\infty} \beta^t u\left(c_t\right)$. As will become clear with the procedure below, this different assumption amounts to a different discount factor by the social planner. In our benchmark formulation, with positive population growth, we give more weight to future utility than in the alternative formulation. Quantitatively, the population growth rates tend to be small in advanced economies, and the impact of the alternative assumptions on the outcome is unlikely to be significant.

The resource constraint with growth becomes

$$C_t = F(K_t, A_t L_t) + (1 - \delta) K_t - K_{t+1}.$$

Defining variables in "per efficiency units of labor", $\tilde{x}_t = X_t / (A_t L_t)$, we can follow similar steps as those in Section 3.2 to write the resource constraint as

$$\tilde{c}_t = f(\tilde{k}_t) + (1 - \delta)\tilde{k}_t - (1 + \gamma)(1 + n)\tilde{k}_{t+1}. \tag{4.18}$$

We further assume that the instantaneous utility is CRRA, $u(c_t) = c_t^{1-\sigma}/(1-\sigma)$, $\sigma > 0$ and $\sigma \neq 1$. Appendix 4.A.2 shows that this function (and the logarithmic function, which should be seen as the case $\sigma \to 1$) is the only form that is consistent with balanced growth. Normalizing $A_0 = 1$ and $L_0 = 1$, we can rewrite the objective function as

$$\sum_{t=0}^{\infty} \beta^t L_t \frac{c_t^{1-\sigma}}{1-\sigma} = \sum_{t=0}^{\infty} \beta^t (1+n)^t \frac{((1+\gamma)^t \tilde{c}_t)^{1-\sigma}}{1-\sigma}$$
$$= \sum_{t=0}^{\infty} \tilde{\beta}^t \frac{\tilde{c}_t^{1-\sigma}}{1-\sigma}, \tag{4.19}$$

where now we use the adjusted discount factor $\tilde{\beta}$:

$$\tilde{\beta} \equiv \beta(1+n)(1+\gamma)^{1-\sigma}. \tag{4.20}$$

The maximization problem for the social planner can then be written as maximizing (4.19) subject to the resource constraint (4.18) and the non-negativity constraints $\tilde{c}_t \geq 0$ and $k_{t+1} \geq 0$. The problem can be solved using standard procedures, and the Euler equation can be derived as:

$$(1+n)(1+\gamma)(\tilde{c}_t)^{-\sigma} = \tilde{\beta}(\tilde{c}_{t+1})^{-\sigma} \left[f'(\tilde{k}_{t+1}) + (1-\delta) \right].$$

Note that, given the definition of $\tilde{\beta}$ in (4.20), the term $(1+n)$ drops out when the Euler equation is written with the original discount factor β. Therefore, the social planner's intertemporal allocation is not affected by the population growth under our utilitarian assumption.

Along the balanced growth path, each variable grows at a constant rate. From our normalized model above, the balanced growth path can be found by imposing steady-state conditions in the $(\tilde{c}_t, \tilde{k}_{t+1})$ variables. This procedure reduces the Euler equation to

$$(1+n)(1+\gamma) = \tilde{\beta} \left[f'(\bar{\tilde{k}}) + (1-\delta) \right].$$

By setting $n = \gamma = 0$, we can check that the above equations coincide with the ones derived under no growth. If we assume a Cobb-Douglas production function, replacing $f'(\bar{\tilde{k}}) = \alpha \bar{\tilde{k}}^{\alpha-1}$ in the equation above, we can deliver $\bar{\tilde{k}}$ as a function of the relevant parameters of the model.

The question of whether or not the economy will converge to its balanced growth path from arbitrary initial conditions will be dealt with in the next section.

4.4 Recursive methods

In the previous section, we solved the maximization problem searching for a sequence of real numbers $\{x_{t+1}^*\}_{t=0}^{\infty}$ that achieves the highest value of the objective function. This involved finding a solution to an infinite sequence of equations (e.g., a difference equation). It is conceptually useful to break down this high-dimensional problem into a sequence of similar, but smaller problems, which all are tied to each other: "recursive" refers to this tie. This principle is at the core of the recursive method known as *dynamic programming*, introduced

by Richard E. Bellman in the 1950s. A key difference with sequential methods is that the solution to the optimization problem will now be a *function* rather than a sequence of numbers. In what follows, we present the idea behind recursive methods and how they can be used; we discuss the precise theoretical links between the sequential approach and the functional approach briefly in Section 4.4.3 below.

4.4.1 Dynamic programming and the Bellman equation

An implicit assumption in the sequential formulation was that the whole path of x_{t+1} was chosen in the initial period. The key to dynamic programming is to think of dynamic decisions as being made not once-and-for-all but period by period instead. In other words, the value of x_{t+1} is decided in period t rather than at date 0. A key question is whether the two formulations are identical. They will be, as long as the problem at hand is **stationary**. This is the case whenever the structure of the maximization problem that a decision maker faces is identical in nature at every point in time. To make ideas more concrete, let's revisit the infinite-horizon version of the problem (P1).

$$V(x_0) \equiv \max_{\{x_{t+1}\in\Gamma(x_t)\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \mathcal{F}(x_t, x_{t+1}) \tag{4.21}$$

where $\Gamma(x_t)$ represents the feasible choice set for x_{t+1} given x_t. In the expression, the "value function" $V(x_0)$ represents the value of the objective function at the optimum given the initial condition x_0. This high-dimensional problem can be broken down and re-written as

$$\begin{aligned}
V(x_0) &= \max_{\{x_{t+1}\in\Gamma(x_t)\}_{t=0}^{\infty}} \left\{ \mathcal{F}(x_0, x_1) + \beta \sum_{t=1}^{\infty} \beta^{t-1} \mathcal{F}(x_t, x_{t+1}) \right\} \\
&= \max_{x_1\in\Gamma(x_0)} \left\{ \mathcal{F}(x_0, x_1) + \beta \left[\max_{\{x_{t+1}\in\Gamma(x_t)\}_{t=1}^{\infty}} \sum_{t=1}^{\infty} \beta^{t-1} \mathcal{F}(x_t, x_{t+1}) \right] \right\} \\
&= \max_{x_1\in\Gamma(x_0)} \left\{ \mathcal{F}(x_0, x_1) + \beta \underbrace{\left[\max_{\{x_{t+2}\in\Gamma(x_{t+1})\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^{t} \mathcal{F}(x_{t+1}, x_{t+2}) \right]}_{\equiv V(x_1)} \right\}
\end{aligned}$$

The simple mathematical idea that $\max_{x,y} f(x,y) = \max_y \{\max_x f(x,y)\}$, provided that each of the max operators is well-defined, allows us to maximize "in steps." To do so, first notice that the problem in squared brackets (in the last row) has the same structure as equation (4.21), but with a different initial condition. In other words, it represents the value of the objective function attained by an agent that chooses the optimal sequence of x_{t+1} given the initial condition x_1. Because the time horizon is the same, and neither the instantaneous objective function nor the feasible set Γ change over time, then it must be the case that the problem in squared brackets is equal to $V(x_1)$. This implies that we can write

$$V(x_0) = \max_{x_1\in\Gamma(x_0)} \left\{ \mathcal{F}(x_0, x_1) + \beta V(x_1) \right\}.$$

The two representations are identical as long as the problem is stationary. Intuitively, a dynamic problem is stationary if the problem at t and at $t+1$ look the same, so one can capture all relevant information for the decision maker in a way that does not involve time. Not all problems are stationary. For example, consider the finite-horizon version of the neoclassical growth model. In it, agents care about how many periods are left when choosing investment. The decision problem changes as the terminal period approaches. One can still describe the problem recursively: break it up into a current choice between consumption and investment and all future choices as implicitly captured by a value function. The difference is that the value function will depend on time, V_t.

With infinitely many periods, the remaining horizon is the same at every t. For different values of the initial capital stock, the choices differ, but do not depend on the time period in which the choice is made. This means that a decision at any point in time does not depend on anything but the level of capital at that point in time.

In general, the predetermined, payoff-relevant information for the consumer is—as was mentioned earlier in this chapter—called a *state variable*. The state variable for the planner in the neoclassical growth model is the current stock of capital k_t. In our generic example, it is simply given by x_t. When transforming a sequential problem into a recursive one, it is important to choose the state variables appropriately so that the resulting problem is indeed recursive.

When the problem is stationary, decisions take a stationary form:

$$x_{t+1} = g(x_t).$$

In particular, the function determining how future capital depends on the current state does not vary with time. The function $g(\cdot)$ is known as the **decision rule** or **policy function**.

Given an arbitrary initial condition x_t, we can write the recursive problem as

$$V(x_t) = \max_{x_{t+1} \in \Gamma(x_t)} \{\mathcal{F}(x_t, x_{t+1}) + \beta V(x_{t+1})\}.$$

This is the dynamic-programming formulation. The derivation was completed for a given value of x_t on the left-hand side of the equation. On the right-hand side, however, we need to know V evaluated at any value for x_{t+1} in order to be able to perform the maximization. Going forward, we now change our notation and use x to denote current values and, adding a prime, x', to denote next period's values. This change of notation is conceptually important: in a stationary dynamic program, "t" actually has no place as time is no longer of essence. Above, it was used to connect the sequence problem to the recursive one; the only remaining need is to distinguish today's x from tomorrow's x'. Thus, we write our dynamic-programming equation as the quest for a $V(x)$ satisfying

$$V(x) = \max_{x' \in \Gamma(x)} \{\mathcal{F}(x, x') + \beta V(x')\} \tag{4.22}$$

for all values of x. This equation is called the "Bellman equation." It is a *functional equation*: the unknown is a function $V(x)$; i.e., it needs to satisfy equation (4.22) for all values of the argument of the function, x. It can be intuitively interpreted as follows: the discounted lifetime value of our representative agent's objective function is equal to the instantaneous

value $\mathcal{F}(x, x')$ received today, plus the discounted value enjoyed from tomorrow into the future, $\beta V(x')$, all assuming optimal choices.

We use the function g alluded to above to denote the arg max in the functional equation:

$$g(x) = \arg \max_{x' \in \Gamma(x)} \{\mathcal{F}(x, x') + \beta V(x')\}$$

for all x, or the decision rule for x': $x' = g(x)$. This notation presumes that a maximum exists and it is unique. Otherwise, g would either not exist or not be a function and, rather, a correspondence. Note that the following must hold for all x by definition.

$$V(x) = \mathcal{F}(x, g(x)) + \beta V(g(x)).$$

4.4.2 Writing a problem recursively

Sometimes we are presented with the sequential formulation of a problem, and it is necessary to 'translate it' into a recursive formulation. We can do this for several reasons. Perhaps the most important one is conceptual: finding a recursive formulation offers an added understanding of the problem at hand. That is because when a problem has been formulated recursively, we know what behavior will depend on: the state variable(s) of the recursive formulation, and on nothing else. In practice, this often means that we can obtain insights into how a problem can be solved. Another reason for studying the recursive problem is that it is associated with convenient computational methods as will be described in Chapter 10.

Consider the infinite-horizon version of the neoclassical growth model, which corresponds to (P3) with $T = \infty$. We first need to make sure that the problem is indeed stationary. That is, that the decision-maker faces the same type of problem every period given the same value of the state variables. That brings us to the question of what the relevant state variables ought to be.

The standard NGM In the NGM, the state variable is given by the stock of capital at the outset of the period, k. In other words, $x = k$ in this model. The value function depends on k, implying that we can write $V(k)$. The control variable is $c = f(k) + (1 - \delta)k - k'$, where we already used the notation that next period variables are denoted with primes (e.g. k_{t+1} is written as k'). The recursive formulation of the neoclassical growth model is

$$V(k) = \max_{k' \in \Gamma(k)} \{u(f(k) + (1 - \delta)k - k') + \beta V(k')\}, \tag{P4}$$

with $\Gamma(k) = [0, f(k) + (1 - \delta)k]$, for all k. While this problem is relatively straightforward to write in recursive form, this is not always the case, as we see with the next two examples.

The periodic NGM:

Consider a small deviation from the problem above, where the level of technology oscillates deterministically between two values A_h and A_l, where $A_h > A_l$. In particular, period t output equals $A_h f(k_t)$ if t is even and $A_l f(k_t)$ if t is odd. The resource constraint is $c_t + k_{t+1} = A_t f(k_t) + (1 - \delta)k_t$, as before. At first sight, this problem does not look stationary because the level of productivity is changing over time. However,

it is changing deterministically and predictably. We can write the problem recursively exploiting the periodicity in the evolution of TFP. To do so, we need to incorporate an additional state variable: whether we are in an even or odd period, which in turn corresponds to an h or l value for A_t. We can therefore write

$$V_h(k) = \max_{k' \in \Gamma_h(k)} \{u(A_h f(k) + (1-\delta)k - k') + \beta V_l(k')\}$$

$$V_l(k) = \max_{k' \in \Gamma_l(k)} \{u(A_l f(k) + (1-\delta)k - k') + \beta V_h(k')\}$$

with $\Gamma_i(k) = [0, A_i f(k) + (1-\delta)k - k']$ for $i \in \{h, l\}$, for all k. Clearly, given the same values of these two states, the decision maker would always select the same k'. So we were able to write the problem recursively with the appropriate choice of state variables.

The delayed depreciation NGM: The objective and resource constraints are standard,

$$\begin{aligned} \max_{\{c_t\}_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t u(c_t) \\ \text{s.t.} \quad & c_t + i_t = F(k_t), \end{aligned}$$

but capital depreciates fully in two periods, and does not depreciate at all before that, so that the law of motion for capital given a sequence of investments $\{i_t\}_{t=0}^{\infty}$ is given by

$$k_t = i_{t-1} + i_{t-2}.$$

In the first period, $k_0 = i_{-1} + i_{-2}$, with two initial conditions i_{-1} and i_{-2}. The recursive formulation for this problem (with "primes" denoting consecutive periods) becomes

$$\begin{aligned} V(i', i) = \quad & \max_{c, i''} \{u(c) + V(i'', i')\} \\ \text{s.t.} \quad & c = f(i' + i) - i''. \end{aligned}$$

Notice that there are two state variables in this problem. That is unavoidable here; there is no way of summarizing what one needs to know at a point in time with only one state variable. Both i_{-1} and i_{-2} are natural state variables: they are predetermined, they affect utility and decision making, and neither is redundant.

4.4.3 Properties of the value function

We have argued that there are two different ways of solving a dynamic maximization problem: (i) the sequential method, where the task is to find a sequence (that maximizes the objective function), and (ii) dynamic programming, where the task is to find a function solving the recursively stated problem, i.e., solving a functional equation. A formal mathematical statement that these two methods are equivalent is beyond the scope of this text; Stokey and Lucas (1989) has all the details. For each of the two approaches, it is important to specify a space within which one looks for a solution—a space of "allowable" sequences and a space of allowable functions, respectively. For dynamic programming, one most conveniently restrict attention to bounded, continuous functions V. For this choice to work out,

one needs assumptions on primitives ($\mathcal{F}$, Γ, and β in our generic formulation); minimum conditions involve (i) continuity of $\mathcal{F}$, (ii) non-emptiness, continuity, and compactness of Γ, and (iii) $\beta < 1$.[13] Under these assumptions, the functional equation defined by the dynamic program has a unique solution (it is the value function what is unique; the policy function can be a correspondence, unless we place further restrictions on $\mathcal{F}$ and Γ.) Moreover, for any given value of the state, it allows us to construct a sequence of choices corresponding to solutions to the sequentially stated problem. Conversely, solutions in the sequence space can be used to construct a value function solving the dynamic programming problem.

We now state some key properties of the dynamic programming problem and briefly comment on them. For proofs, see Stokey and Lucas (1989). You may have studied the necessary ingredients into the proofs in preparatory math, in particular the fact that the dynamic program defines a *contraction mapping*.

1. It is possible to find $V(x)$ by an iterative process. The procedure is as follows.

   i. Select any initial $V_0(x)$ function. One example is $V_0(x) = 0 \ \forall x$.

   ii. Define a sequence of functions as follows: for all x,

   $$V_{n+1}(x) = \max_{x' \in \Gamma(k)} \mathcal{F}(x, x') + \beta V_n(x')$$

   for $n = 0, 1, 2, \ldots$.

   Then the following is true:

   (i) the resulting sequence $\{V_j(x)\}_{j=0}^{\infty}$ converges to $V(x)$, i.e., to the function that solves the dynamic program;

   (ii) the distance to the solution V gets smaller and smaller at a constant rate: $\|V_{n+1} - V\| \leq \beta \|V_n - V\|$, where $\|f\|$ denotes a distance between functions.[14]

   Notice that the particular initial guess $V_0 = 0$ delivers an interpretation of V_n: it represents present-value utility for an economy consisting of n periods. Clearly, V_n is an addition of "utils" over time which involves optimal choice of x' in all periods, including the last one, V_1.

   Feature (ii) is rather remarkable: it holds no matter what the initial guess is and it is therefore very useful in practical applications: for finding a solution numerically. In particular, if one can bound the initial error by some value $\bar{\epsilon} \equiv \|V_0 - V\|$ and then generate the sequence of functions, one knows that after n iterations the error is at most $\beta^n \bar{\epsilon}$. Finally, (ii) is extremely useful in allowing us to prove properties of the value function, such as those we state next.

[13] Some maximization problems involve discrete choice; we could, for example, restrict the choice of capital to belong to a finite set of values. The conditions on primitives then become correspondingly weaker: $\mathcal{F}$ has to be bounded and Γ nonempty.

[14] The distance is a norm; in practice the sup-norm is used, i.e., the largest difference across all values of the argument of the function, i.e., $\|f\| \equiv \sup_x f(x)$.

2. Assume that $\mathcal{F}$ is strictly increasing in its first argument and that Γ is monotone: if $x \leq \tilde{x}$, then $\Gamma(x) \subseteq \Gamma(\tilde{x})$. Then V is strictly increasing.

3. Assume that $\mathcal{F}$ is strictly concave (in its two arguments jointly) and that $\Gamma(x)$ is convex in the following sense: if $x' \in \Gamma(x)$ and $\tilde{x}' \in \Gamma(\tilde{x})$, then $\theta x' + (1-\theta)\tilde{x}' \in \Gamma(\theta x + (1-\theta)\tilde{x})$. Then V is strictly concave and the policy unique (i.e., g is a well-defined function).

4. Assume that $\mathcal{F}$ and Γ satisfy the properties in the previous statement and, in addition, that $\mathcal{F}$ is continuously differentiable. Then for any x where the choice is interior, V is differentiable, i.e., $V'(x)$ exists.

5. Assume that $\mathcal{F}$ and Γ satisfy all the conditions above. Then, the policy function $g(x)$ is strictly increasing.

Very briefly, property 1 above follows from the Bellman equation being a contraction mapping.[15] Properties 2 and 3 are proved using property 1 as follows: take a V_0 with the desired characteristic (e.g., concavity), show that if V_n has the desired characteristic then so does V_{n+1}, and finally argue that the limit V^* inherits the characteristic at least weakly (e.g., V^* is concave). Strictness of the characteristic (e.g., V^* is strictly concave) follows by applying the argument again, with the limit V^* inserted on the right-hand side of the Bellman equation. Property 4 is also possible to prove relatively straightforwardly but note that it applies only in the case where V is concave. Property 5 follows from the first-order condition

$$-\mathcal{F}_2(x, x') = \beta V'(x').$$

The left-hand side of this equality is clearly increasing in x', since $\mathcal{F}(x, x')$ is strictly concave in its second argument, and the right-hand side is strictly decreasing in x', since $V(x)$ is strictly concave under the stated assumptions. Furthermore, since the right-hand side is independent of x but the left-hand side is decreasing in x, the optimal choice of x' is increasing in x.

4.4.4 Solving for the value function

As for the sequentially formulated maximization problem, analytical solutions are only available in very special cases. When they do exist, there are again a few different methods available to find them. One is a "guess and verify method," also known as the "method of undetermined coefficients." This method, however, requires an insightful initial guess for $V(x)$. When the initial guess is unavailable, we can use the iterative process described above, starting from $V_0(x) = 0$, quite similarly to how we proceeded to solve the finite-horizon model backwards.

We again illustrate these solution methods using the neoclassical growth model, (P4). Further details can be found in Chapter 10.

Example 4.7 *Consider the economy from Example 4.5. The corresponding Bellman equation is*

$$V(k) = \max_{k' \geq 0} \left\{ \log\left(Ak^{\alpha} - k'\right) + \beta V(k') \right\}.$$

[15] A contraction mapping obtains for the Bellman equation so long as $\beta < 1$.

Let the initial guess of the value function be $V_0(k) = 0$. Then,

$$V_1(k) = \max_{k' \geq 0} \{\log [Ak^\alpha - k']\}. \tag{4.23}$$

The right-hand side is maximized by taking $k' = 0$, yielding $V_1(k) = \log A + \alpha \log k$. Using this in the next iteration, we obtain

$$V_2(k) = \max_{k' \geq 0} \{\log [Ak^\alpha - k'] + \beta [\log A + \alpha \log k']\}.$$

The first-order condition delivers

$$\frac{1}{Ak^\alpha - k'} = \frac{\beta\alpha}{k'} \Rightarrow k' = \frac{\alpha\beta A k^\alpha}{1 + \alpha\beta}.$$

We can interpret the resulting expression for k' as the rule that determines how much it would be optimal to save if we were at period $T-1$ in the finite horizon model. We can substitute this into $V_2(k)$ to yield

$$\begin{aligned} V_2(k) &= \log \left[Ak^\alpha - \frac{\alpha\beta A k^\alpha}{1 + \alpha\beta}\right] + \beta \left[\log A + \alpha \log \frac{\alpha\beta A k^\alpha}{1 + \alpha\beta}\right] \\ &= \underbrace{\log \left(A - \frac{\alpha\beta A}{1 + \alpha\beta}\right) + \beta \log A + \alpha\beta \log \frac{\alpha\beta A}{1 + \alpha\beta}}_{=a_2} + \underbrace{\left(\alpha + \alpha^2\beta\right)}_{=b_2} \log k. \end{aligned}$$

The same procedure can be used to obtain a $V_3(k)$, and so on. This procedure would make the sequence of value functions converge to $V(k)$.

We can also, at this point, guess and verify a functional form for V. Grouping terms in V_2, we can see that it takes the form $V_n(k) = a_n + b_n \log k$ for all n. Therefore, we may already guess that the function to which this sequence is converging has to be of the form:

$$V(k) = a + b \log k.$$

In order to determine the corresponding parameters a, b, we take first-order conditions and find that $k' = \dfrac{\beta b}{1 + \beta b} A k^\alpha$. Inserting this into the right-hand side of the Bellman equation

$$RHS \equiv \max_{k' \geq 0} \{\log (Ak^\alpha - k') + \beta (a + b \log k')\}.$$

and equating the resulting expression to our guess delivers a system of two equations in two unknowns, a and b. The solutions will be

$$b = \frac{\alpha}{1 - \alpha\beta} \quad \textit{and} \quad a = \frac{1}{1 - \beta} \frac{1}{1 - \alpha\beta} \left[\log A + \log (1 - \alpha\beta)^{1 - \alpha\beta} + \log (\alpha\beta)^{\alpha\beta}\right].$$

The resulting decision rule is exactly the same policy rule obtained using sequential methods,

$$k' = \alpha\beta A k^\alpha.$$

4.4.5 The functional Euler equation

In the sequentially formulated maximization problem, the Euler equation turned out to be a crucial part of characterizing the solution. With the recursive strategy, an Euler equation can be derived as well. Consider the Bellman equation for a general instantaneous objective function $\mathcal{F}(x, x')$:

$$V(x) = \max_{x' \in \Gamma(x)} \{\mathcal{F}(x, x') + \beta V(x')\}.$$

Under suitable assumptions, this problem delivers the policy function $x' = g(x)$. Hence,

$$V(x) = \mathcal{F}(x,\ g(x)) + \beta V(g(x)). \tag{4.24}$$

Assuming an interior solution, $g(x)$ satisfies the first-order condition

$$\mathcal{F}_2(x,\ x') + \beta V'(x') = 0.$$

Evaluated at the optimum, i.e., at $x' = g(x)$,

$$\mathcal{F}_2(x,\ g(x)) + \beta V'(g(x)) = 0.$$

In contrast to how we derived the sequential formulation, we now need to know the derivative of the value function $V'(\cdot)$ in order to solve for the policy rule. Even though it is not possible, in general, to write $V(x)$ in terms of primitives, we can find its derivative. Using the equation (4.24) above, one can differentiate both sides with respect to x (recall that the equation holds for all x and, again under some assumptions stated earlier, is differentiable). We obtain

$$V'(x) = \mathcal{F}_1(x,\ g(x)) + \underbrace{g'(x)\Big[\mathcal{F}_2(x,\ g(x)) + \beta V'(g(x))\Big]}_{\text{indirect effect through optimal choice of } x'},$$

where "$g'(x)$" represents the derivative of the policy function. The differentiability of g was not established above, and it is harder to prove, but it is actually not required in order to derive the result that just will follow. Namely, from the first-order condition, the argument in brackets in the equation just stated is zero and hence

$$V'(x) = \mathcal{F}_1(x,\ g(x)),$$

which again holds for all values of x. The indirect effect thus disappears: this is an application of a general result known as the **Envelope Theorem**.

Updating, we know that $V'(g(x)) = \mathcal{F}_1\Big(g(x),\ g(g(x))\Big)$ also has to hold. The first-order condition can now be rewritten as follows:

$$\mathcal{F}_2(x,\ g(x)) + \beta \mathcal{F}_1\Big(g(x),\ g(g(x))\Big) = 0\ \forall x. \tag{4.25}$$

This is the Euler equation stated as a functional equation: it does not contain the unknowns x_t, x_{t+1}, and x_{t+2} but, rather, has to hold for all x. Recall our previous Euler equation formulation

$$\mathcal{F}_2(x_t,\ x_{t+1}) + \beta \mathcal{F}_1(x_{t+1},\ x_{t+2}) = 0, \forall t,$$

where the unknown was the sequence $\{x_t\}_{t=1}^{\infty}$. Now instead, the unknown is the function g. That is, under the recursive formulation, the Euler equation turned into a functional equation: the **functional Euler equation**.

Solving directly for policy The previous discussion suggests that a third way of searching for a solution to the dynamic problem is to consider the functional Euler equation, and solve it for the function g. We have previously seen that we can (i) look for sequences solving a nonlinear difference equation plus a transversality condition; or (ii) we can solve a Bellman (functional) equation for a value function.

The functional Euler equation offers another way to solve directly for behavior, thus bypassing the value function. Its key feature is that it expressed an intertemporal tradeoff. Here, the recursive approach provides some extra structure relative to the sequential Euler equation: it tells us that the optimal sequence of capital stocks needs to be connected using a stationary function. Mathematically, unlike the Bellman equation, the functional Euler equation is not a contraction mapping. It is also only a necessary condition on the optimal policy function g and, viewed in isolation, can allow multiple solutions. Only one of these solutions, however, is the policy function that solves the right-hand side of the Bellman equation. It will become clear below how, at least locally, multiple solutions are possible.

The functional Euler equation approach is often used in practice in solving dynamic problems numerically. We now show an example for which an analytical solution exists.

Example 4.8 *Consider the model used in Example 4.7. With full depreciation* $\mathcal{F}(k,k') = u(f(k)-k')$. *Then, the respective derivatives are:*

$$\begin{aligned}\mathcal{F}_1(k,k') &= u'(f(k)-k')f'(k)\\ \mathcal{F}_2(k,k') &= -u'(f(k)-k').\end{aligned}$$

In the particular parametric example, and replacing $k' = g(k)$, *equation (4.25) becomes:*

$$\frac{1}{Ak^\alpha - g(k)} - \frac{\beta\alpha A\left(g(k)\right)^{\alpha-1}}{A\left(g(k)\right)^\alpha - g\left(g(k)\right)} = 0, \forall k.$$

This is a functional equation in $g(k)$. *Guess that* $g(k) = sAk^\alpha$, *i.e., saving is a constant fraction of output. Substituting this guess into functional Euler equation delivers*

$$\frac{1}{(1-s)Ak^\alpha} = \frac{\alpha\beta A\left(sAk^\alpha\right)^{\alpha-1}}{A\left(sAk^\alpha\right)^\alpha - sA\left(sAk^\alpha\right)^\alpha}.$$

As can be seen, k *cancels out, and the remaining equation can be solved for* s. *Collecting terms and factoring out* s, *we obtain*

$$s = \alpha\beta.$$

Thus, $\alpha\beta Ak^\alpha$ *satisfies the functional Euler equation for all values of* k. *It is, moreover, the same answer that we arrived at in Example 4.7.*

4.4.6 Dynamics in the optimizing neoclassical growth model

We now apply recursive methods to characterize the solution to the planning problem involving the standard neoclassical growth model. We could also apply our methods to the case of consumer saving under a constant wage and interest rate, but we leave that application for the reader. It may be useful to first note that there is a unique maximum attainable

level of the capital stock, k_u. That is, if the capital stock starts out below (or at) k_u, it cannot take a higher value than k_u in the future. This result follows because for any k, $f(k) + (1 - \delta)k$ is the highest feasible saving k', because this implies zero consumption. Hence, $f(k_u) + (1 - \delta)k_u = k_u$ defines k_u and we can restrict attention to a search for value and policy functions over the closed and bounded set $[0, k_u]$. There are two solutions to this equation: a strictly positive k_u and 0; we discard the latter.

We know from the functional Euler equation (4.25) that, for all $k \in [0, k_u]$, $g(k)$ satisfies

$$u'\left(f(k) + (1 - \delta)k - g(k)\right) = \beta u'\left(f(g(k)) + (1 - \delta)g(k) - g(g(k))\right)\left(f'(g(k)) + 1 - \delta\right) \tag{4.26}$$

In particular, at a steady state $\bar{k}$, $g(\bar{k}) = \bar{k}$, which—given that the argument of u' is non-zero—allows us to write the steady-state condition as

$$1 = \beta(f'(\bar{k}) + 1 - \delta), \tag{4.27}$$

which is identical to (4.10) above. Now we can proceed to show global convergence to the steady state.

We use the property that under the standard assumptions the value function $V(k)$ is strictly concave.[16] This property implies that we can write

$$\left[V'(k) - V'(g(k))\right]\left[k - g(k)\right] \leq 0 \quad \forall k \in [0, k_u].$$

To see that the inequality holds, note that whenever $g(k) > k$, the expression contained in the left-most bracket is positive, whereas the right-most bracket contains a negative quantity, and vice versa.

Using the envelope theorem, we know that

$$V'(k) = u'\left(f(k) + (1 - \delta)k - g(k)\right)\left(f'(k) + 1 - \delta\right).$$

From the first-order condition in the Bellman equation, we obtain

$$V'(g(k)) = u'\left(f(k) + (1 - \delta)k - g(k)\right)\frac{1}{\beta}.$$

Inserting these two expressions into the left-most bracket of the inequality and factorizing u', we obtain

$$u'(f(k) - g(k))\left[f'(k) + 1 - \delta - \frac{1}{\beta}\right]\left[k - g(k)\right] \leq 0 \quad \forall k \in [0, k_u]. \tag{4.28}$$

When $k = \bar{k}$, from (4.27), $f'(k) + 1 - \delta = 1/\beta$ and $g(k) = k$ holds. When $k > \bar{k}$, then $f'(k) + 1 - \delta < 1/\beta$, and thus (4.28) implies $g(k) < k$, and capital decreases over time. When $k < \bar{k}$, capital increases. This result implies, given that we also know that g is increasing (property 5 above), that the system is globally stable and converges monotonically to $\bar{k}$.

[16] This requires us to show that $u(f(k) + (1 - \delta)k - k')$ is concave in (k, k'). This can be accomplished by computing the Hessian and showing that it is negative definite.

Approximating the policy function Most often, it is not possible to solve for the policy function in analytical form. In such cases, we can approximate the solution locally with a linear function,

$$g(k) \sim a_0 + a_1 k$$

where a_0 and a_1 are the coefficients that we need to solve for, around the steady-state value of capital. We need two equations to solve for our two unknowns, but since $g(k) = k = \bar{k}$ in steady state we obtain one condition as

$$a_0 = \bar{k}(1 - a_1),$$

where $\bar{k}$ solves (4.10). To obtain a_1, we use a procedure that parallels linearization techniques discussed in Chapter 3, but is less cumbersome because we exploit the fact that the functional Euler equation must hold for all k. Thus, we can simply differentiate equation (4.26) with respect to k, i.e., take the derivative of the left-hand side, LHS, and set it equal to the derivative of the right-hand side, RHS. We obtain, with the obvious notation, $LHS_k(k) = RHS_k(k)$, which must hold for all k. Evaluating this expression at steady state allows us to obtain an equation that provides a second condition relating a_1 to a_0. Hence,

$$LHS_k(\bar{k}) = u''(\bar{c})[f'(\bar{k}) + 1 - \delta - \underbrace{g'(\bar{k})}_{=a_1}]$$

$$RHS_k(\bar{k}) = \beta u''(\bar{c})\Big[(f'(\bar{k}) + 1 - \delta)a_1 - \underbrace{g'(\bar{k})g'(\bar{k})}_{=a_1^2}\Big]\Big[f'(\bar{k}) + 1 - \delta\Big] + \beta u'(\bar{c})f''(\bar{k})a_1.$$

where we used the fact that, under our guess, $g(g(k)) = a_0 + a_1[a_0 + a_1 k]$; furthermore, $\bar{c}$ is defined as $f(\bar{k}) - \delta\bar{k}$. Setting $LHS_k(\bar{k}) = RHS_k(\bar{k})$ and simplifying (recall $f'(\bar{k}) + 1 - \delta = 1/\beta$) delivers

$$u''(\bar{c})\left[1 - \beta a_1\right] = u''(\bar{c})\left[a_1 - \beta a_1^2\right] + \beta^2 u'(\bar{c})f''(\bar{k})a_1, \tag{4.29}$$

which can be used to solve for a_1 given $\bar{k}$. Clearly, a_1 satisfies a second-order polynomial equation. It is straightforward to show that this equation has two real solutions, of which one is strictly between zero and one, and the other is strictly greater than one. Thus, locally, $g(k)$ can be approximated by two functions. This was alluded to above and it is not a surprise, given that the functional Euler equation is only a necessary condition. A function g which solves this equation also has to attain the maximum on the right-hand side of the Bellman equation. Of the two functions obtained here, only one has that property: the one with an a_1 between zero and one. It is the one corresponding to a slope less than one in (k, k') space, thus giving monotone convergence to the steady state. Concretely, by denoting $\Theta \equiv \beta^2 u' f''/u'' > 0$, the solution of (4.29) is

$$a_1 = \frac{1 + \beta + \Theta - \sqrt{(1 + \beta + \Theta)^2 - 4\beta}}{2\beta}. \tag{4.30}$$

It can easily be checked that $a_1 \in (0, 1]$, a_1 is decreasing in Θ, and $a_1 = 1$ when $\Theta = 0$ and $a_1 \to 0$ as $\Theta \to \infty$.

Using this procedure, it is possible to also obtain higher-order approximations to the policy function g around the steady state. To obtain the second-order Taylor approximation, simply differentiate the functional equation once more with respect to k, which delivers g'' at the steady state: it can be solved for as a function of g and g', which were previously solved-out. Further differentiations will, successively, give us any desired higher-order approximations.[17]

Calibration We discussed the concept of calibration in Chapter 3: the adoption of specific functional forms and parameter values with the purpose of generating quantitative predictions. We applied it in that chapter to gauge the Solow model's quantitative predictions for the speed of convergence to a steady state. The NGM developed in this chapter, instead, differs in its convergence properties: the saving rate is not constant away from steady state but is instead endogenous and, as we have seen, a function of the current level of the capital stock. How strong the dependence on capital is depends on the utility function. More precisely, equation (4.30) determines a_1, the slope of the saving function as it crosses the steady-state line: a slope of zero implies convergence in one period (infinite speed) and a slope of 1 implies no (infinitely slow) convergence. The remaining model parameters can be selected as in Chapter 3 so our only question here is: what is a_1, or rather, what is the value of Θ at the steady state? Recall that $\Theta = \beta^2 u' f''/u''$, $\bar{c} = f(\bar{k}) - \delta\bar{k}$, and $f'(\bar{k}) = 1/\beta - (1-\delta)$. Furthermore, suppose that $u(c) = (c^{1-\sigma} - 1)/(1-\sigma)$. Then it can be shown that

$$\Theta = -\frac{\beta^2 f''(\bar{k})(f(\bar{k}) - \delta\bar{k})}{\sigma}.$$

Because $f''(\bar{k})$ is negative, Θ is positive and decreasing in σ. When $\sigma = 0$, the utility function is linear, and in this case $\Theta \to \infty$. The above result implies $a_1 \to 0$ in this case, implying k' does not depend on k. Convergence, in this case, is immediate. The opposite extreme is when $\sigma \to \infty$, $\Theta \to 0$ and $a_1 \to 1$. In this case, the utility function is "infinitely curved" (Leontief utility). When σ is very large, consumers are extremely unwilling to change consumption over time and convergence is very slow.

Therefore, σ is an important parameter in determining the speed of convergence. What do empirical studies of consumption suggest as an appropriate value of σ? Using aggregate consumption data, Hall (1988a) estimates that $1/\sigma = 0.1$. Attanasio and Weber (1993) and Attanasio and Weber (1995) use micro data instead and finds that $1/\sigma \in [0.3, 0.8]$. In many applied macroeconomic studies using the CRRA function, however, there are also stochastic components, and under uncertainty, σ is also equal to the coefficient of relative risk aversion (uncertainty is discussed in Chapter 7 below) and this coefficient is often estimated to be above 1. Perhaps for this reason, much of the applied literature focuses on $\sigma = 1$ (log utility) or $\sigma = 2$, though rarely much higher than one.

Further assume that the production function $f(k)$ is a power function: $f(k) = k^\alpha$. This formulation implies the aggregate production function is of the Cobb-Douglas form. In this case, Θ can be solved as

$$\Theta = \frac{\beta^2(1-\alpha)}{\sigma}\left(\frac{1}{\beta} - 1 + \delta\right)\left[\frac{1}{\alpha}\left(\frac{1}{\beta} - 1 + \delta\right) - \delta\right],$$

[17] Approximations based on Taylor approximations, as those carried out here, of course require differentiability of the function. Thus, we approximate g, which is endogenous, with that proviso.

which is decreasing in α and increasing in δ. Convergence is slower when the production function is closer to linear and the depreciation rate is small. This qualitative property is shared with the Solow model.

4.5 Concluding remarks

In this chapter, we showed a set of tools that are useful for solving standard dynamic optimization models used in macroeconomics. We discussed the sequential formulation, where the aim is to choose the best sequence of allocations (quantities) that maximize lifetime utility, starting with a finite horizon economy and then moving to an infinite horizon economy, highlighting the mathematical complications that arise. We then showed how to use dynamic programming methods to write the optimization problem recursively, where the key is to solve for policy functions determining optimal allocations. In practice, all these methods are used. We spent less time on motivating specific functional forms and, throughout, only used two examples: a pure consumption-saving problem under price-taking and an optimizing neoclassical growth model. Neither of these cases allowed valued leisure or other, richer optimization problems. However, the methods we introduce here are straightforwardly extended to other, richer contexts and will indeed be used over and over in the rest of the text.

Chapter 5

Dynamic competitive equilibrium

The introductory chapter looking at long-run data, Chapter 2, argued that a certain view on production and technical change was important for understanding our macroeconomic history. It also argued, informally at least, that a market economy with certain features was another important component of the overall picture: firms and private households making decisions in their self interest. The most commonly used framework in economics is a market setting with perfect competition—indeed, it is the setting the earlier chapter alluded to—and in the present chapter we will develop such theory carefully in the context of a dynamic economy.

Like many of our assumptions, perfect competition should be viewed as a useful approximation rather than an exact description of our economy. Indeed, most firms have a degree of market power and in some industries market power is critical for understanding how the industry works. However, a macroeconomic approach naturally starts from a benchmark case that is (i) not too far from a description of "average" behavior and (ii) allows tractable analysis, while still capturing the key features of markets. These features involve how private incentives—of consumers and firms—jointly steer production and consumption through a price mechanism. Market scarcity—expressed as a demand that exceeds supply—will lead to higher prices, lowering demand and/or raising supply. For example, the level of investment, a key macroeconomic variable, is affected by the interest rate, which is a relative price between goods today and goods in the future: consumers are willing to forsake goods today by lending more to investors the higher is the interest rate; and investors thus invest until the interest rate they have to pay on a loan no longer balances the productive returns on the investment. This market mechanism, moreover, is at play not just for traditional consumption and investment goods and services but for the development of new ideas and technologies that are key to long-run development.

How do agents decide what to buy or sell? They solve an optimization problem that is—in a macroeconomic context—often dynamic, as we have seen in Chapter 4. The solution to the optimization problem provides the quantities of the objects that the agent would like to buy and/or sell at given prices; for example, a household could sell labor services and get paid a wage (the price for labor). Therefore, an important component of a competitive equilibrium is the characterization of individual demands and supplies for goods, services and

assets, *given* the corresponding prices.[1] Since in the economy there are many agents who buy and sell the traded objects, the derivation of economy-wide demands and supplies requires the aggregation of demands and supplies over all agents. The core framework described in this chapter makes this procedure simpler by assuming that all consumers are identical and that all firms are identical—we use the "representative consumer" and the "representative firm" as our key constructs. A large part of the current macroeconomic research literature of course focuses on consumer and firm heterogeneity, which will be amply discussed later in the text.

In summary, a perfectly competitive equilibrium occurs when two conditions are met: agents' choices are optimal taking the prices as given and markets clear, i.e., demand equals supply, for all goods and services. In this chapter, we will also see that the market allocation delivers a Pareto optimal outcome. With a representative consumer, this amounts to verifying that, in the perfectly competitive equilibrium, the consumer's utility is maximized given all technological constraints. When we depart from this setting, and when monopoly elements and other "frictions" are introduced, markets need no longer deliver optimal outcomes. We look at such instances in some detail in Chapter 6 but they will then appear again and again in the applied chapters later.

5.1 Different equilibrium concepts

When considering our macroeconomic models, it is very helpful to be precise in formulating equilibrium concepts. A model economy, first, has some descriptive "deep" features, such as a set of agents (consumers, firms, a government, etc.) along with their objectives (such as utility functions) and relevant constraints (time available, technologies, etc.). Second, what the set of traded goods and services is—along with their prices—must be specified. The equilibrium concept then lists a set of conditions that need to be fulfilled: (i) agents maximizing their objectives subject to their constraints (e.g., consumers maximizing utility subject to their budget constraints and firms maximizing profits taking their technological possibilities into account) and (ii) resource feasibility: that what is consumed is also produced, often expressed as "demand equals supply," for all traded goods and services. The definition of an equilibrium thus organizes this information in the form of a set of conditions that need to be fulfilled. In this chapter, we will formulate equilibrium definitions in a rather mathematical way, i.e., with a minimum of words. Thus, we will try to refrain from conditions such as "taking prices as given" in a maximization problem; rather, if a maximization problem is specified, it must list the choice variables, and it then follows that any variables that are not listed as choices (such as prices in a perfectly competitive equilibrium) must, then, be taken as given.

Formulating equilibria in dynamic models, moreover, involves a choice regarding how time is dealt with. We will, in particular, consider three different definitions of competitive equilibria in turn:

1. Arrow-Debreu equilibrium in which trades occur at date 0,

[1] The notion that agents take prices as given allows us to analyze the economy without game-theoretic elements: in a given agent's decision problem, the behavior of others appears in the form of (endogenous) constants.

2. sequential equilibrium in which trades occur period by period, and

3. recursive equilibrium in which prices and decisions are expressed as functions of the economy's state variables.

In the first two of these, an equilibrium consists of sequences (of quantities and prices) indexed by time, i.e., a specific values for each variable at each point in time. The third concept, in contrast, formulates the equilibrium as a set of functions of state variables (such as a capital stock or level of asset holdings). For many economies, one can define the competitive equilibrium in any of these three ways and they all give rise to the same allocations. Which one is chosen in a given application depends on the purpose, but, broadly speaking, the Arrow-Debreu equilibrium is the version that aligns most closely with microeconomic theory while being rather abstract as a concept. Sequential equilibria are defined more in accordance with how we think events actually play out in reality. A recursive equilibrium has the sequential trading structure but is expressed in terms of functions rather than sequences because it uses dynamic programming methods. Whereas equilibria based on sequences typically are defined given specific values of the state variables (such as the initial capital stock), recursive equilibria are not, i.e., they apply to all values of the state variables.

Throughout, we will focus on dynamic models with an infinite time horizon. Finite-horizon economies can of course be studied too; they should be seen as special cases of what we study here. As pointed out above, infinite-horizon economies are useful because there, time plays a less central role: the remaining time horizon is always the same. Occasionally (later in the text), we focus on one-period (where the "time horizon" plays no role) or two-period economies, but later in the text. In Sections 5.2–5.4 we study infinitely-lived dynasties and go through the different equilibrium definitions. In the final part of the chapter, Section 5.5, we study overlapping-generations (OG) economies, where time goes on forever but all individuals live a finite number of periods.

5.2 Arrow-Debreu equilibrium

In an Arrow-Debreu equilibrium, all trades take place at time zero. One way to think about this assumption is that every agent signs a contract at time zero with other agents. The contract specifies the quantities of the traded objects that the agent will deliver or receive at any future time t from other agents at the prices specified. This market structure is called an Arrow-Debreu or date-0 market. It differs from the sequential market structure where trades for goods take place at all times t, not only at time zero. This different market structure will be considered in the other two equilibrium concepts.

It is perhaps worth emphasizing that signed contracts are fully enforceable, that is, promises made in a contract are always fulfilled. If some of the promises are not enforceable, we are in an environment in which markets are incomplete. We will consider economies with incomplete markets in later chapters.

The best way to illustrate the concept of an Arrow-Debreu equilibrium is through the application to specific economies. We start with an endowment economy.

5.2.1 An endowment economy

The first example is an economy in which production is exogenously determined. There is a continuum of consumers, indexed by $i \in [0,1]$. In every period t each consumer produces $y_{i,t} \in \mathbb{R}_+$ of a single consumption good that cannot be stored for future consumption. The model does not specify how production takes place (for example with the input of labor) and the exogenous production is often referred to as endowment. An economy without the specification of a production technology is sometimes called an *exchange economy*, since the only economic activity that agents undertake, besides consumption, is the trade of the endowments. But, effectively, the exchange economy can be seen as a special case of a production economy. Finally, the consumer's utility from any given consumption path $\{c_{i,t}\}_{t=0}^{\infty}$ is

$$\sum_{t=0}^{\infty} \beta^t u\left(c_{i,t}\right). \tag{5.1}$$

Consumers' preferences are all the same (they have the same u and β).

What is traded here is goods at different points in time: buying one unit of good at time t means buying a contract that gives ownership of a unit of good at that point in time. Its price is denoted by p_t. Thus, p_t/p_0 represents the units of consumption goods delivered at time 0 that are needed to buy 1 unit of the consumption good delivered at time t. It is customary to normalize the price at time zero to 1 so that the good at time 0 becomes the numéraire. Then p_t is the price of a time-t good in terms of time-0 consumption goods.

Given the price p_t for $t = 0, 1, ...$, the total value of the consumer's endowments is given by $\sum_{t=0}^{\infty} p_t y_{i,t}$ and the value of the consumer's total expenditures is $\sum_{t=0}^{\infty} p_t c_{i,t}$. The budget constraint then requires that the value of expenditures not be larger than the value of the endowments. In practice, since we always use strictly increasing utility functions, consumers will always use up all the resources and we will therefore impose this constraint with equality. Thus, we have

$$\sum_{t=0}^{\infty} p_t c_{i,t} = \sum_{t=0}^{\infty} p_t y_{i,t}. \tag{5.2}$$

We will define the equilibrium focusing on the mathematical conditions that must be satisfied, as opposed to their economic interpretation. The equilibrium definition provides a roadmap for solving for an equilibrium—we convert the equilibrium conditions into a set of equations that allow us to solve the model.[2]

Definition 1 *An **Arrow-Debreu competitive equilibrium** is a set of sequences $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$, for each $i \in [0,1]$, and $\{p_t\}_{t=0}^{\infty}$ such that*

1. *for each i, $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$ solves*

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \quad \text{subject to} \quad \sum_{t=0}^{\infty} p_t c_t = \sum_{t=0}^{\infty} p_t y_{i,t};$$

[2]Here and in what follows, we omit any requirements that quantities and prices be non-negative. After we solve for an equilibrium we can verify that any such requirements are satisfied. Also, we do not specify the mathematical spaces to which the sequences must belong; for infinite sequences, this involves advanced concepts, which is why we omit them.

2. *for each* t, $\int_0^1 c_{i,t}^\star di = \int_0^1 y_{i,t} di$.

Equilibrium characterization Let us now characterize equilibrium. We will gradually introduce more detail for our assumptions regarding the utility function and endowment sequences. To characterize the optimal decision of a consumer (point 1 in the above definition), we write the Langragian for the consumer problem. We then take the first-order conditions as shown in Chapter 4. This gives us

$$\beta^t u'(c_{i,t}) = \lambda_i p_t,$$

where λ_i denotes the Lagrange multiplier for the (lifetime) budget constraint, equation (5.2), and the prime denotes the derivative of the utility function. The multiplier is the shadow value of lifetime wealth which could differ across agents if their lifetime endowments differ. However, the multiplier does not depend on time. If the utility function is strictly concave, which we will assume here, its derivative is strictly decreasing in consumption. Thus, the above condition determines, uniquely, the optimal consumption for agent i.

If we eliminate the multiplier using the optimality conditions at time zero and at a generic time t, for the same consumer, we obtain the price

$$p_t = \beta^t \frac{u'(c_{i,t})}{u'(c_{i,0})}. \tag{5.3}$$

Remember that p_0 has been normalized to 1, which explains why it does not appear in this expression. Since the price p_t is the same for all agents, this condition tells us that the ratio of marginal utilities at different times are the same across consumers. This equation states that the relative price of consumption at time t in terms of time zero consumption has to equal the marginal rate of substitution between these two goods. This is the ratio between the present value of the marginal utility of consumption at time t and the marginal utility of consumption at time zero.

To gain further intuition, we now consider the special case in which the utility function takes the logarithmic form, that is, $u(\cdot) = \log(\cdot)$. With this special utility, the above condition becomes

$$p_t = \beta^t \frac{c_{i,0}}{c_{i,t}}.$$

Since this condition must hold for all agents, we can use this condition to find equilibrium prices. Multiply both sides of the equation by $c_{i,t}$ and sum across all i. We obtain $p_t \int_0^1 c_{i,t} di = \beta^t \int_0^1 c_{i,0} di$. Using market clearing, i.e., $\int_0^1 c_{i,t} di = \int_0^1 y_{i,t} di = Y_t$, we obtain

$$p_t = \beta^t \frac{Y_0}{Y_t}.$$

Thus, we see that prices decline over time because of discounting but also that periods with lower total endowments have higher prices. This is because marginal utility is strictly decreasing: lower total resources make each unit of consumption more valuable on the margin.

From the Euler equation, we see that in this economy all consumers experience the same consumption growth. More specifically, if we take any two consumers, i and i', we have

$$\frac{c_{i,t+1}}{c_{i,t}} = \frac{c_{i',t+1}}{c_{i',t}} = \frac{Y_{t+1}}{Y_t} = \beta \frac{p_t}{p_{t+1}}.$$

This condition is valid for any pattern of the individual endowments. Clearly, the growth rate of individual consumption is only determined by the aggregate endowment Y_t. Thus, individual consumption could be less volatile than individual income (consumption smoothing). To see this more clearly, suppose for a moment that individual endowments change over time but that the aggregate endowment, $Y_t = \bar{Y}$, does not. Therefore, $\int_0^1 y_{i,t} di = \bar{Y}$. In equilibrium it must be that $\int_0^1 c_{i,t} di = \bar{Y}$. In other words, aggregate consumption is constant. Since the consumption growth of all agents is the same, this implies that individual consumption must also be constant for all $t = 0, 1, 2, \ldots$. Therefore, in this special case with a constant aggregate endowment, but not necessarily constant individual endowments, the model features perfect consumption smoothing.

Lastly, the level of consumption for each consumer can be derived from the individual budget constraint. By combining the Euler equations for $t = 0, 1, 2, ...$ and $p_0 = 1$ we arrive at $p_t c_{i,t} = \beta^t c_{i,0}$. The budget constraint then becomes

$$c_{i,0} \sum_{t=0}^{\infty} \beta^t = Y_0 \sum_{t=0}^{\infty} \beta^t \frac{y_{i,t}}{Y_t} \quad \Rightarrow \quad c_{i,0} = (1-\beta) Y_0 \sum_{t=0}^{\infty} \beta^t \frac{y_{i,t}}{Y_t}.$$

Clearly, the level of consumption of individual i, as measured by its level at time 0, is a fraction $1-\beta$ of the individual's present-value income. This income is a function of endowments where (i) endowments further into the future obtain lower weights due to discounting and (ii) endowments in periods where aggregate income is high obtain a lower weight. The latter is true since resources in periods with high aggregate income are given lower value by consumers: their marginal utilities are lower then than in other periods. Thus, two consumers with the same average endowments can have different total wealth: the wealthy consumer is the one whose endowments are high when others' endowments are low.

Clearly, since consumption growth is identical for all consumers, we see that individual i's share of aggregate consumption, and aggregate resources, is constant over time:

$$c_{i,t} = \theta_i C_t = \theta_i Y_t, \quad \text{with} \quad \theta_i = (1-\beta) \sum_{t=0}^{\infty} \beta^t \frac{y_{i,t}}{Y_t}.$$

In the special case in which agents have exactly the same endowment ($y_{i,t} = Y_t$), the share is 1 for all agents which effectively means that they consume their own endowment. We are then in the case of a *representative consumer*.

If the utility function is not logarithmic but maintains the balanced-growth form $(c^{1-\sigma} - 1)/(1-\sigma)$ (with $\sigma > 0$ and $\sigma \neq 1$; recall that $\sigma \to 1$ should be interpreted as $\log c$), then it is still possible to solve the model. It is easy to verify, using the consumer's Euler equation, that individuals' consumption growth rates will all be identical: the gross rates will equal $(\beta p_t / p_{t+1})^{\frac{1}{\sigma}}$. From this it follows that $p_t = \beta^t (Y_0/Y_t)^{\sigma}$: again, resources are more valuable in times with lower endowments, and the more so the higher is the curvature of the utility function, as measured by σ.

The endowment framework studied here is useful in many contexts. One of these is asset pricing, covered in Chapter 16. The idea there is that any asset can be thought of as a stream of payments, "dividends," such as the endowment sequences described here, so the price of the asset is then the total market value of these endowments. Prices of endowments

at each time period are straightforward to compute in endowment economies, including in cases with uncertainty; we will look at uncertainty in Chapter 7.

Finally, in many macroeconomic applications, the assumption of a representative agent is used. It makes sense as a special case of the above, when u is strictly concave: then, given equal endowments for all agents, the consumption choices must all be the same. Such an equilibrium is often defined more compactly as

Definition 2 *An **Arrow-Debreu competitive equilibrium** is a set of sequences* $\{c_t^\star\}_{t=0}^\infty$ *and* $\{p_t\}_{t=0}^\infty$ *such that*

1. $\{c_t^\star\}_{t=0}^\infty$ *solves*

$$\max_{\{c_t\}_{t=0}^\infty} \sum_{t=0}^\infty \beta^t u(c_t) \quad \textit{subject to} \quad \sum_{t=0}^\infty p_t c_t = \sum_{t=0}^\infty p_t y_t;$$

2. $c_t^\star = y_t$.

This equilibrium definition is mathematically precise but the economic context—that of many consumers, making the same choices—is only written "between the lines."

5.2.2 A production economy with labor

We now consider an economy where there is production but with only one factor input: labor. The economy is populated by a continuum of households, each supplying working hours $\ell_{i,t}$ to the market. However, working is costly for the household as it reduces utility. Given the consumption path $\{c_{i,t}\}_{t=0}^\infty$ and working hours $\{\ell_{i,t}\}_{t=0}^\infty$, the household's utility is

$$\sum_{t=0}^\infty \beta^t \left[u\left(c_{i,t}\right) - v(\ell_{i,t})\right]. \tag{5.4}$$

The function $v(\ell_{i,t})$ is the disutility from working. It is increasing and convex in $\ell_{i,t}$. We allow households to differ in their efficiency units of labor which we denote by e_i. What this means is that, if household i works $\ell_{i,t}$ hours, its contribution to the economy's input of labor (in efficiency units) is $e_i \ell_{i,t}$. We can think of e_i as labor skills.

In this economy there is also a representative firm that produces consumption goods with the production function

$$Y_t = A_t L_t, \tag{5.5}$$

where A_t could be time varying but not stochastic, and L_t is the effective input of labor used in production.

In a competitive economy, the presumption is also that there are many firms. However, if these firms all have the same objective (profit maximization) and the same technology available, then they will face identical maximization problems. For this reason, we will use the concept of a *representative firm*. It thus hires labor from households paying the wage rate w_t per effective unit of labor in terms of the consumption good at that time. Therefore, if the household supplies one efficiency unit of labor at time t, it receives w_t units of consumption goods at that time, which translates to $p_t w_t$ units in terms of time-0 good.

The representative firm's objective is to maximize profits:

$$\max_{\{L_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \Big\{ p_t A_t L_t - p_t w_t L_t \Big\}. \tag{5.6}$$

Clearly, L_t only appears in the time-t term, so this problem reduces to an infinite sequence of static problems $\max_{L_t}(A_t L_t - w_t L_t)$; the within-period relative price w_t is the price that will clear the time-t market. Unlike for the concept of the representative consumer, whose optimal choice is unique, the optimal firm choice is—under constant returns to scale and price-taking—not unique in equilibrium: the firms is indifferent as to what scale to operate at.[3]

Given prices p_t and $p_t w_t$ and labor supplies $\ell_{i,t}$, for $t = 0, 1, ...$, the value of lifetime income for household i is $\sum_{t=0}^{\infty} p_t w_t e_i \ell_{i,t}$ and the value of its expenditures is $\sum_{t=0}^{\infty} p_t c_{i,t}$. The budget constraint requires that the value of expenditures not be larger than the value of incomes, that is,

$$\sum_{t=0}^{\infty} p_t c_{i,t} = \sum_{t=0}^{\infty} p_t w_t e_i \ell_{i,t}. \tag{5.7}$$

The following is then the compact definition of our equilibrium.

Definition 3 *An **Arrow-Debreu competitive equilibrium** is a set of sequences* $\left\{c_{i,t}^{\star}\right\}_{t=0}^{\infty}$ *and* $\left\{\ell_{i,t}^{\star}\right\}_{t=0}^{\infty}$*, for each* $i \in [0,1]$*,* $\{L_t^{\star}\}_{t=0}^{\infty}$*,* $\{p_t\}_{t=0}^{\infty}$ *and* $\{w_t\}_{t=0}^{\infty}$ *such that*

1. *for each* i*,* $\left\{c_{i,t}^{\star}\right\}_{t=0}^{\infty}$ *and* $\left\{\ell_{i,t}^{\star}\right\}_{t=0}^{\infty}$ *solve*

$$\max_{\{c_t,\ell_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \left[u(c_t) - v(\ell_t)\right] \quad \text{subject to} \quad \sum_{t=0}^{\infty} p_t c_t = \sum_{t=0}^{\infty} p_t w_t e_i \ell_t;$$

2. $\{L_t^{\star}\}_{t=0}^{\infty}$ *solves*

$$\max_{\{L_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \Big\{ p_t A_t L_t - p_t w_t L_t \Big\};$$

3. *for each* t*,* $\int_0^1 e_i \ell_{i,t}^{\star} di = L_t^{\star}$ *and* $\int_0^1 c_{i,t}^{\star} di = A_t L_t^{\star}$.

The first two points say that the allocation is what agents (households and firms) choose optimally to maximize their respective objective functions.[4] The third item defines the market clearing condition for labor (the aggregate supply of labor from households must be equal to the aggregate demand from firms) and for the goods markets (the aggregate demand of goods from households must be equal to the aggregate supply, which corresponds to production).

[3] In this case, an equilibrium price for labor, w_t, has to equal A_t, in which case the firm is indifferent as to its choice of L_t; if $A_t > w_t$, there is no optimal scale (profits rise without bound with scale), and if $A_t < w_t$, the firm's unique maximizer is $L_t = 0$.

[4] For the firm, we have written it so that the firm chooses labor and production at all times. Since the problem is not fundamentally dynamic—L_t can be chosen independently of labor chosen at times other than t—we could alternatively had firms operating production at single dates only.

Equilibrium characterization The optimality conditions for the consumer's problem (again derived from the Lagrangian as shown in Chapter 4) are

$$\beta^t u'(c_{i,t}) = \lambda_i p_t, \tag{5.8}$$

and

$$v'(\ell_{i,t}) = e_i w_t u'(c_t), \tag{5.9}$$

where λ_i is the Lagrange multiplier for the (lifetime) budget constraint, equation (5.7).

We can eliminate the multiplier using the optimality conditions for an agent i, at time t and at time $t+1$, and obtain

$$\frac{p_{t+1}}{p_t} = \frac{\beta u'(c_{i,t+1})}{u'(c_{i,t})}. \tag{5.10}$$

As in the case of the endowment economy we just studied, the relative price of tomorrow's consumption in terms of today's consumption—the inverse of the (gross) real interest rate—equals the marginal rate of substitution between consumption at t and $t+1$. With the normalization $p_0 = 1$, the equilibrium price at any time t is

$$p_t = \frac{\beta^t u'(c_{i,t})}{u'(c_{i,0})}.$$

The firm solves (5.6) and the first-order condition requires that, for an interior solution—where the firm is indifferent between using more or less labor—we need to have

$$A_t = w_t.$$

If A_t were to exceed w_t, the firm's maximization problem would have no solution: the higher is L_t, the better. If w_t were to exceed A_t the firm would shut down (choose zero labor). Thus, for an equilibrium to exist, the relative price w_t needs to adjust to be equal to A_t. When it has, the firm's size is indeterminate: all $L_t \geq 0$ deliver zero profits. This is a standard result under constant returns to scale and will apply also when the firm has more inputs (such as labor an capital, as in the neoclassical growth model). Our notation may suggest that there is one firm that produces the economy's entire output but, really, we think of a large number of firms solving the same problem under perfect competition, and the equilibrium will then determine total production but not which firm produces how much.

We now again consider the special case in which the utility function takes the logarithmic form, that is, $u(\cdot) = \log(\cdot)$. With this utility equation (5.10) can be written as

$$\frac{c_{i,t+1}}{c_{i,t}} = \frac{\beta p_t}{p_{t+1}}.$$

Since prices are the same for all agents, this implies that households experience the same growth in consumption. Thus, also in this case we have that individual consumption is a constant share of aggregate output, that is, $c_{i,t} = \theta_i Y_t$. However, output is not exogenous but it depends on the endogenous input of labor L_t. This is determined by the aggregation of individual labor supplies which are determined by the first-order condition (5.9). Using the

log specification of the utility function and the fact that the wage rate is given by $w_t = A_t$, the condition can be rewritten as

$$v'(\ell_{i,t})c_{i,t} = e_i A_t.$$

Next we use the property that individual consumption is a fixed share of aggregate output $c_{i,t} = \theta_i Y_t$. Substituting in the first-order condition we obtain

$$v'(\ell_{i,t})\theta_i Y_t = e_i A_t.$$

Let us normalize skills so that $\int_0^1 e_i di = 1$. Given that skills are constant over time, different consumers have different total resources to spend in exact proportion to e_i, aside from differences in how much they work. However, we see that if we conjecture $\theta_i = e_i$, then the above condition implies that all households will supply the same labor $\ell_{i,t} = L_t$, confirming that total resources available to spend are simply proportional to e_i.

To see that the lifetime budget (5.7) is satisfied, use $c_{i,t} = \theta_i Y_t$, $w_t = A_t$, and $\ell_{i,t} = L_t$ in the budget constraint to obtain

$$\theta_i \sum_{t=0}^{\infty} p_t A_t L_t = e_i \sum_{t=0}^{\infty} p_t A_t L_t.$$

This is satisfied if $\theta_i = e_i$, confirming our guess.

Because high-skilled households earn higher incomes, they will enjoy higher consumption. High-skilled households, however, supply the same amount of labor as do low-skilled households. This is a consequence of income and substitution effects offsetting each other. To see this, note that an extra unit of work effort earns $e_i w_t$, which increases with the wage and the skill of the worker. This generates a substitution effect in the direction of working more. However, since—without working more—the household earns higher consumption the higher is the wage, it experiences a lower marginal utility of consumption, which leads it to wish to work less (choose lower effort/higher leisure). With the utility function here, the two effects exactly cancel. These features play out even more clearly in a static optimization problem where the consumer has labor income only and spends it on consumption, $c = w\ell$, with the utility function $\log c - v(\ell)$: the choice of ℓ will not depend on w. This feature was alluded to in Chapter 2, where we argued that balanced growth with constant labor supply restricts us to a specific class of utility functions, to which the current example belongs.

Finally, imagine that individuals, in addition to being able to work, had independent ("asset") income. In particular, individual i would be endowed with $a_{i,0}$ at time zero, with $\int_0^1 a_{i,0} di = 0$. This assumption implies that, if some agents have positive asset holdings, other agents must have negative holdings; thus, the resource constraints remain unchanged. Now, agents with different asset holdings will work different amounts because assets generate an income effect but not a substitution effect: higher asset holdings, by making the individual richer, will induce lower labor effort, as discussed above. Consider the simple static case just described: if the budget constraint reads $c_i = e_i A \ell_i + a_i$, individual labor supply will, using the first-order condition, be given by $v'(\ell_i) = a_i A/(e_i A \ell_i + a_i)$. Here, ℓ_i will depend nonlinearly on a_i (so long as v' is not a constant). Hence, total labor supply, along with total consumption, will depend on the distribution of assets: they will depend on wealth inequality. Consumption growth would still be equalized across agents, but total production will also depend on inequality since total consumption does.

5.2.3 The neoclassical growth economy

We extend the economy considered in the previous subsection by adding capital to production. This is essentially the neoclassical growth model with endogenous supply of labor. The accumulation of capital makes the problem more complex because what is produced in the future depends on the capital that is accumulated today. In most cases, a full analytical solution will not be available. Nevertheless, we can define the conditions that an equilibrium must satisfy.

The production function now takes the form

$$Y_t = A_t K_t^{\alpha} L_t^{1-\alpha}, \tag{5.11}$$

where A_t is productivity, K_t is the input of capital and L_t is the input of labor.

We assume that capital is accumulated by households who then rent it to firms, similarly to labor. Therefore, in addition to the price for goods and labor, we now have a price for rental capital. The representative firm solves the profit maximization problem

$$\max_{\{K_t, L_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \Big\{ p_t A_t K_t^{\alpha} L_t^{1-\alpha} - p_t r_t K_t - p_t w_t L_t \Big\}, \tag{5.12}$$

where r_t denotes the rental price of capital paid to households expressed in time-t consumption. As in the previous sub-section, the optimal choice of the firm for period t does not affect the optimal choices in other periods; we can equivalently state the problem as a sequence of sub-problems: for each t, $\max_{K_t, L_t} A_t K_t^{\alpha} L_t^{1-\alpha} - r_t K_t - w_t L_t$.

In this economy, production is used for consumption, C_t, and for investment, I_t:

$$Y_t = C_t + I_t.$$

Given the initial capital held by a household, $k_{i,0}$, and individual i's investment $\iota_{i,t}$, for $t = 0, 1, ...$, the individual stock of capital evolves according to

$$k_{i,t+1} = (1 - \delta) k_{i,t} + \iota_{i,t}. \tag{5.13}$$

Given p_t, r_t, w_t, investment $\iota_{i,t}$, and labor supply $\ell_{i,t}$, for $t = 0, 1, ..$, the value of lifetime income is $\sum_{t=0}^{\infty} p_t (r_t k_{i,t} + w_t e_i \ell_{i,t})$ and the value of expenditures for consumption and investment is $\sum_{t=0}^{\infty} p_t (c_{i,t} + \iota_{i,t})$. The budget constraint requires that the value of expenditures be equal to the value of incomes, that is,

$$\sum_{t=0}^{\infty} p_t (c_{i,t} + \iota_{i,t}) = \sum_{t=0}^{\infty} p_t (r_t k_{i,t} + w_t e_i \ell_{i,t}). \tag{5.14}$$

The compact equilibrium definition reads as follows.

Definition 4 *An **Arrow-Debreu competitive equilibrium** is a set of sequences* $\left\{c_{i,t}^{\star}\right\}_{t=0}^{\infty}$, $\left\{\iota_{i,t}^{\star}\right\}_{t=0}^{\infty}$, *and* $\left\{\ell_{i,t}^{\star}\right\}_{t=0}^{\infty}$, *for each* $i \in [0,1]$, $\left\{L_t^{\star}\right\}_{t=0}^{\infty}$, $\left\{K_{t+1}^{\star}\right\}_{t=0}^{\infty}$, $\left\{p_t\right\}_{t=0}^{\infty}$, $\left\{p_t w_t\right\}_{t=0}^{\infty}$ *and* $\left\{p_t r_t\right\}_{t=0}^{\infty}$ *such that*

1. *for each* i, $\{c^\star_{i,t}\}_{t=0}^\infty$, $\{\iota^\star_{i,t}\}_{t=0}^\infty$, *and* $\{\ell^\star_{i,t}\}_{t=0}^\infty$ *solve*

$$\max_{\{c_t,\iota_t,\ell_t\}_{t=0}^\infty} \sum_{t=0}^\infty \beta^t \left[u(c_t) - v(\ell_t)\right] \quad \textit{subject to} \quad \sum_{t=0}^\infty p_t(c_t + \iota_t) = \sum_{t=0}^\infty p_t(r_t k_t + w_t e_i \ell_t)$$

$$\textit{where, } \forall t, \quad k_{t+1} = (1-\delta)k_t + \iota_t \textit{ with } k_0 = k_{i,0},$$

2. $\{L^\star_t\}_{t=0}^\infty$ *and* $\{K^\star_t\}_{t=0}^\infty$, *where* $K^\star_0 = \int_0^1 k^\star_{i,0} di$, *with* $k^\star_{i,0} = k_{i,0}$, *solve*

$$\max_{\{L_t,K_t\}_{t=0}^\infty} \sum_{t=0}^\infty p_t \Big\{ A_t K_t^\alpha L_t^{1-\alpha} - r_t K_t - w_t L_t \Big\},$$

3. $\int_0^1 e_i \ell^\star_{i,t} di = L^\star_t$, $\int_0^1 k^\star_{i,t} di = K^\star_t$, *and* $\int_0^1 (c^\star_{i,t} + \iota^\star_{i,t}) di = A_t (K^\star_t)^\alpha (L^\star_t)^{1-\alpha}$ *for all* t.

Note that the $k_{i,0}$s are exogenous and that K_0 therefore is not included as an equilibrium object: we only include as equilibrium objects those that are endogenously determined. Also note, again, that since firms rent capital from consumers, their profit maximization problem is not dynamic (and we could equivalently have one firm per date). It is also possible to define an equilibrium where firms buy and own capital. A firm would then purchase a unit of the good at t and use it as an investment good and then use it, together with labor at $t+1$. That way, the firm would have a dynamic problem, making profits in period $t+1$ and on that, in equilibrium, would be just large enough to offset the costs of investment at t because of constant returns to scale.

Equilibrium characterization The optimality conditions for consumption, labor, and investment are derived by formulating the Lagrangian and taking derivatives. For consumption and labor we have, as in the previous economy,

$$\beta^t u'(c_{i,t}) = \lambda_i p_t, \tag{5.15}$$

$$v'(\ell_{i,t}) = e_i w_t u'(c_{i,t}), \tag{5.16}$$

where λ_i is the Lagrange multiplier for the (lifetime) budget constraint, equation (5.14).

Since the optimality condition for consumption is the same as in the previous economy, we can again show that all households experience the same consumption growth independently of their skills e_i and their initial wealth $k_{i,0}$, so long as $u(c)$ is a power function, which we will assume. This also implies that individual consumption is a share θ_i of aggregate consumption, that is, $c_{i,t} = \theta_i C_t$. What determines θ_i are, as before, the skills e_i and the initial wealth $k_{i,0}$, along with how prices develop over time.

The optimality condition for investment is new and it takes the form

$$p_t = p_{t+1} r_{t+1} + (1-\delta) p_{t+1}. \tag{5.17}$$

The condition has a simple interpretation. If we buy one unit of capital today, it will cost us p_t. Next period, however, we can rent it to the representative firm earning the rental rate $p_{t+1} r_{t+1}$. In addition, we still have the non-depreciated capital $1-\delta$, which is worth

$(1-\delta)p_{t+1}$. Thus, the left-hand-side is the cost of investing and the right-hand-side is the gross return, both expressed in terms of time-0 consumption. The equality simply says that the cost of investing must be equal to its return.[5] Note that if we divide both sides of equation (5.17) by p_{t+1} we find $p_t/p_{t+1} = 1 + r_{t+1} - \delta$ so the price of consuming in date t in terms of foregone consumption in $t+1$ is given by the gross return on capital.

The firm's problem now has two optimality conditions,

$$\begin{aligned} r_t &= \alpha A_t K_t^{\alpha-1} L_t^{1-\alpha}, \\ w_t &= (1-\alpha) A_t K_t^{\alpha} L_t^{-\alpha}. \end{aligned}$$

The rental rate of capital and the wage rate are, respectively, the marginal productivities of capital and labor. Again, for general pairs (r_t, w_t) these two conditions will not be met at the same time (in which a firm would make infinite profits or shut down): these prices will, however, adjust so that they are, and the firm is then indifferent as to the scale of the operation. However, the capital-labor ratio is pinned down uniquely.

Let us now specialize the utility function further: $u(c) = \log c$. Then, using the first-order condition for consumption, equation (5.15), at t and $t+1$, we obtain

$$c_{i,t+1} = \beta \frac{p_t}{p_{t+1}} c_{i,t}.$$

Since $c_{i,t} = \theta_i C_t$ for all t, we obtain (equivalently, sum across all i)

$$C_{t+1} = \beta \frac{p_t}{p_{t+1}} C_t,$$

i.e., the Euler equation can also be expressed in terms of aggregate consumption. However, from the discussion at the end of Section 5.2.2 we know that aggregate labor supply will depend on the asset distribution—as given by the distribution of $k_{i,0}$s here—so let us make a simplifying assumption here:

$$k_{i,0} = e_i K_0.$$

This assumption implies if you are rich in assets, you are also rich in productivity. Here, e_i can then be interpreted also as the share of total initial capital K_0 held by agent i. We can also see that, conditional on working the same amounts, agents with different e_is will have total wealth (including the asset holding) levels proportional to e_i. To see this, use $\iota_{i,t} = k_{i,t+1} - (1-\delta)k_{i,t}$ and (5.17) for all t in the individual budget (5.14) to obtain

$$\sum_{t=0}^{\infty} p_t c_{i,t} = \sum_{t=0}^{\infty} p_t w_t e_i \ell_{i,t} + (1 - \delta + r_0) k_{i,0}.$$

Here we see that, if we set $\ell_{i,t} = \ell_t$ independent of i—a guess we will confirm to be correct momentarily—then individual i's total resources (the right-hand side) would be proportional

[5] From the consumer's perspective, one could conceive of prices and rental rates such that the stated equality is an inequality. This would either mean that a maximum does not exist or a corner solution; in equilibrium, neither are possible, so prices have to adjust to ensure equality.

to e_i, since $k_{i,0}$ also is. Since the Euler equations yield $p_{t+1}c_{i,t+1} = \beta p_t c_{i,t}$ for all t and i, we can then insert, simplify, and rewrite the budget as

$$c_{i,t} = e_i \frac{\beta^t(1-\beta)}{p_t}\left(\sum_{t=0}^{\infty} p_t w_t \ell_t + (1-\delta+r_0)K_0\right).$$

We see that θ_i, the individual's share of aggregate consumption, equals e_i.[6] Going back to equation (5.16), we see that the right-hand side becomes independent of i, and hence $\ell_{i,t}$ will be independent of i (recall $u'(c_{i,t}) = 1/c_{i,t}$ with log utility): it will equal L_t.

Given our assumptions, we can now think of there being a "representative consumer," even though consumers differ in endowments, since individuals' decisions all scale with e_i (this representative is, rather, the aggregate of all consumers). We have not solved the model fully yet, however: aggregates remain to be determined. We can collect them as follows:

$$\frac{C_{t+1}}{C_t} = \beta\left(1-\delta+\alpha A_{t+1}K_{t+1}^{\alpha-1}L_{t+1}^{1-\alpha}\right), \tag{5.18}$$

where we have used the aggregate Euler equation, (5.17), and the firm's first-order condition for capital;

$$v'(L_t) = \frac{(1-\alpha)A_t K_t^{\alpha} L_t^{-\alpha}}{C_t}, \tag{5.19}$$

where we have used the consumer's intratemporal first-order condition and the firm's first-order condition for labor; and

$$C_t = A_t K_t^{\alpha} L_t^{1-\alpha} + (1-\delta)K_t - K_{t+1} \tag{5.20}$$

from the resource constraint. Consumption can be eliminated and the system can be written as a difference equation (second-order in K, given L). It does not, in general, have a closed-form solution.[7]

How does the competitive equilibrium allocation compare to the solution to a social planner's problem of the kind we studied in Chapter 4? To explore this, let us suppose that all individuals have the same labor efficiency so $e_i = 1$ for all i. The social planner then wishes to maximize the utility of the representative household

$$\sum_{t=0}^{\infty} \beta^t \left[\log\left(C_t\right) - v\left(L_t\right)\right]$$

subject to the aggregate resource constraint (5.20). If we substitute the constraint into the objective function and take the first order condition for the planner's problem with respect to K_{t+1} we obtain the exact same expression as (5.18). Similarly the first order condition with respect to L_t is exactly the same as (5.19). The equations that characterize the solution to the planner's problem are therefore exactly the same as the equations that the competitive equilibrium must satisfy. As a result, the competitive equilibrium coincides with the choice of the planner. This is an important result that we will return to in Chapter 6.[8]

[6] To see this, note that $c_{i,t}$ has the structure $e_i X_t$ where X_t is independent of i. Integrating across i and using $\int e_i di = 1$ we find that $C_t \equiv \int c_{i,t} di = X_t$.

[7] $\delta = 1$ would deliver a constant saving rate equal to $\alpha\beta$, and L_t would then be given by $v'(L_t)L_t = \frac{1-\alpha}{1-\alpha\beta}$ and not depend on time.

[8] We assumed here that all the individuals have the same labor productivity to simplify the exposition,

5.3 Sequential equilibrium

In the previous sections we defined equilibria in a manner following Arrow-Debreu: all trades are decided on at time zero. In this section, instead, we assume that trades are decided on sequentially over time and deliveries take place either in the same period t or in future periods. Future deliveries are especially relevant for financial contracts. For example, a debt contract is signed at time t, when borrowing occurs, but the repayment arises in one or more future periods.

Since trades arise sequentially, agents face a budget constraint in every period. This, however, does not imply that agents need to consume the whole income earned in the period. They can save by holding assets. In the rest of this section we assume that there is only one asset, denoted by a_t, that pays interest at a net rate r_t. We will define $q_t a_{t+1}$ as the amount saved at t and a_{t+1} as the amount delivered at $t+1$. This means that the real interest rate between t and $t+1$ can be defined from $1 + r_{t+1} = 1/q_t$.[9]

The period resources that the agent does not use for consumption will be used to purchase assets. If instead consumption exceeds the resources available in the period, the agent will borrow. Borrowing means that the value of a_t is negative, in which case the agent will pay an interest (the interests on borrowing and saving are the same). This will become clear in the applications.

5.3.1 The endowment economy

Returning to our endowment economy from Section 5.2.1, let us now assume that agents trade in every period. Agents thus trade an asset that pays the interest rate r_t: agent i enters period t with $a_{i,t}$ units of the asset and receives the endowment $y_{i,t}$. Therefore, the total resources available in the period are $a_{i,t} + y_{i,t}$. These resources are then used in part for consumption, $c_{i,t}$, and in part to purchase new units of the asset, $q_t a_{i,t+1}$. The budget constraint in period t is thus

$$c_{i,t} + q_t a_{i,t+1} = y_{i,t} + a_{i,t}.$$

The agent thus maximizes lifetime utility (5.1) subject to the sequence of budget constraints, one for every period, and the no Ponzi game (nPg) condition introduced in Section 4.3.1. The initial asset holdings, which are exogenous, sum to zero: $\int_0^1 a_{i,0} di = 0$. We have the following.

Definition 5 *A **sequential competitive equilibrium** is a set of sequences* $\left\{c^\star_{i,t}\right\}_{t=0}^{\infty}$ *and* $\left\{a^\star_{i,t+1}\right\}_{t=0}^{\infty}$*, for each* $i \in [0,1]$*, and* $\{q_t\}_{t=0}^{\infty}$ *such that*

but this assumption is not crucial. With heterogeneous skill levels we could show that the competitive equilibrium coincides with the solution to a different social planner's problem in which the social planner maximizes a weighted average of the utilities of the individuals with weights given by e_i.

[9] This convention is common; in an endowment economy, if alternatively a_{t+1} is saving in consumption units at t and $(1 + r_{t+1})a_{t+1}$ is the total return from this saving at $t+1$, then r_0 cannot be determined in equilibrium (or, equivalently, it can be set to any value). However, q_0 is determined.

1. *for each* i, $\left\{c^{\star}_{i,t}\right\}_{t=0}^{\infty}$ *and* $\left\{a^{\star}_{i,t+1}\right\}_{t=0}^{\infty}$ *solve*

$$\max_{\{c_t, a_{t+1}\}_{t=0}^{\infty}} \quad \sum_{t=0}^{\infty} \beta^t u(c_t)$$

subject to

$$c_t + q_t a_{t+1} = a_t + y_{i,t} \quad \forall\, t, \ \textit{with } a_0 = a_{i,0},$$

$$\lim_{t\to\infty} \left(\prod_{s=0}^{t} q_s\right) a_{t+1} \geq 0. \qquad \textit{(nPg condition)}$$

2. *for all* t, $\int_0^1 c^{\star}_{i,t} di = \int_0^1 y_{i,t} di$ *and* $\int_0^1 a^{\star}_{i,t+1} di = 0$.

The last two conditions are the market clearing conditions in goods and financial markets. The first says that the aggregate consumption must be equal to the aggregate quantity of goods available in every period. The second says that aggregate asset holdings must be zero in every period. In this economy, agents borrow and lend to one another. Each transaction therefore represents an increase in one person's assets and an offsetting decrease in another's. When we sum across all the agents, these trades net out to zero.

The equilibrium definition imposes two market clearing conditions, but in fact only one of them is needed due to Walras's Law. If we integrate the household budget constraints across i and impose asset market clearing we arrive at the goods market clearing condition. Hence goods market clearing is insured by asset market clearing. It is also the case that goods market clearing implies the asset market clears.[10]

The optimality conditions can be derived by writing the Lagrangian and taking first-order conditions

$$\beta^t u'(c_{i,t}) = \lambda_{i,t} \tag{5.21}$$

and

$$q_t \lambda_{i,t} = \lambda_{i,t+1}, \tag{5.22}$$

where $\lambda_{i,t}$ is the Lagrange multiplier associated with the budget constraint. Differently from the first-order conditions in the setup with time-zero trading, the multiplier depends on time t. This is because we replaced the lifetime budget constraint (which is one constraint) with the sequence of period budget constraints. Therefore, we have one multiplier associated with each period budget constraint.

We now show that, despite this difference, we obtain the same optimality conditions. Using equation (5.21) at time t and $t+1$ to eliminate the multipliers in equation (5.22) we obtain

$$q_t = \beta \frac{u'(c_{i,t+1})}{u'(c_{i,t})}.$$

[10] If we integrate the budget constraints at time 0 and impose $\int_0^1 a_{i,0} di = 0$ and goods market clearing we find $\int_0^1 a_{i,1} di = 0$. We can then proceed by induction to show that goods market clearing at date 1 implies asset market clearing at date 1 and so on.

This condition must be satisfied at any time t. If we use this condition from time 0 through time $t-1$, we obtain

$$q_0 \times q_1 \times \cdots \times q_{t-1} = \beta \frac{u'(c_{i,1})}{u'(c_{i,0})} \times .. \times \beta \frac{u'(c_{i,t})}{u'(c_{i,t-1})},$$

which can be rewritten more compactly as

$$\prod_{s=0}^{t-1} q_s = \beta^t \frac{u'(c_{i,t})}{u'(c_{i,0})}.$$

The left-hand-side term corresponds to p_t in the Arrow-Debreu equilibrium. Therefore, we obtained the optimality condition (5.3) we derived in the time-zero trade equilibrium. This illustrates that the Arrow-Debreu price p_t is the present value at time zero of one unit of consumption at time t, discounted by the sequence of interest rates up to time t. It also shows that $p_{t-1}/p_t = 1/q_{t-1} = 1 + r_t$, that is, the gross interest rate between $t-1$ and t is the ratio between the corresponding Arrow-Debreu prices: the relative price of consumption goods at $t-1$ in terms of goods at t.

The sequential equilibrium thus delivers the same equations determining quantities and, once translated back into Arrow-Debreu terms, the same prices as well. The literature uses both, guided by what is convenient in different contexts.

It is instructive to connect the budget constraints in the two setups: at first appearance they may not look equivalent, but they are. To see this, we start with the budget constraints at $t = 0$ and $t = 1$:

$$\begin{aligned} c_{i,0} + q_0 a_{i,1} &= a_{i,0} + y_{i,0}, \\ c_{i,1} + q_1 a_{i,2} &= a_{i,1} + y_{i,1}. \end{aligned}$$

Using the first equation to eliminate $a_{i,1}$ in the second equation (or viceversa) we obtain

$$c_{i,0} + q_0 c_{i,1} + q_1 q_0 a_{i,2} = a_{i,0} + y_{i,0} + q_0 y_{i,1}.$$

We use next the budget constraint at $t = 2$ to eliminate $a_{i,2}$, then the budget constraint at $t = 3$ to eliminate $a_{i,3}$, and so on. After T substitutions we obtain

$$\sum_{t=0}^{T} \left(\prod_{s=0}^{t-1} q_s \right) c_{i,t} + \left(\prod_{s=0}^{T-1} q_s \right) a_{i,T+1} = a_{i,0} + \sum_{t=0}^{T} \left(\prod_{s=0}^{t-1} q_s \right) y_{i,t}.$$

We have already shown that $\prod_{s=0}^{t-1} q_s = p_t$. Furthermore, as T converges to infinity, the second term on the left-hand side of the equation converges to something non-negative. Therefore, taking the limit $T \to \infty$ we obtain the lifetime budget constraint

$$\sum_{t=0}^{\infty} p_t c_{i,t} \leq a_{i,0} + \sum_{t=0}^{\infty} p_t y_{i,t}.$$

The budget constraint will be chosen to hold with equality given a strictly increasing utility function; that is, the limit of $\left(\prod_{s=0}^{T-1} q_s \right) a_{i,T+1}$ as t approaches infinity will be zero (a strictly positive amount would violate the TVC discussed in Chapter 4).

It is straightforward to apply a sequential equilibrium concept to other economies; to save space, we will only look at the neoclassical growth model, this time without valued leisure.

5.3.2 The neoclassical growth economy

We can similarly define an equilibrium of the neoclassical growth model in which agents trade period by period. We simply state the compact equilibrium definition here and leave it up to the reader to verify that it is equivalent to that described as an Arrow-Debreu equilibrium: that the allocations coincide and that prices, appropriately defined in comparable terms, do too. We use the case where labor is set exogenously to e_i for agent i, with $\int_0^1 e_i di = 1$ so that aggregate labor supply equals 1.

Definition 6 *A **sequential competitive equilibrium** is a set of sequences $\left\{c^\star_{i,t}\right\}_{t=0}^\infty$ and $\left\{k^\star_{i,t+1}\right\}_{t=0}^\infty$, for each $i \in [0,1]$, and $\{r_t\}_{t=0}^\infty$ and $\{w_t\}_{t=0}^\infty$ such that*

1. *for each i, $\left\{c^\star_{i,t}\right\}_{t=0}^\infty$ and $\left\{k^\star_{i,t+1}\right\}_{t=0}^\infty$ solve*

$$\begin{aligned}\max_{\{c_t,k_{t+1}\}_{t=0}^\infty} \quad & \sum_{t=0}^\infty \beta^t u(c_t) \\ & \text{subject to} \\ & c_t + k_{t+1} = (1-\delta+r_t)k_t + w_t e_i \quad \forall\, t, \text{ with } k_0 = k_{i,0}, \\ & \lim_{t\to\infty} \frac{1}{\prod_{s=0}^t (1-\delta+r_{s+1})} k_{t+1} \geq 0 \qquad \textit{(nPg condition)};\end{aligned}$$

2. *for each t, $(K_t, 1)$ solves $\max_{K,L} A_t K^\alpha L^{1-\alpha} - r_t K - w_t L$, where $K_t = \int_0^1 k^\star_{i,t} di$ (and $k^\star_{i,0} = k_{i,0}$); and*

3. *for each t,*

$$C_t + K_{t+1} = A_t K_t^\alpha + (1-\delta)K_t,$$

where $C_t = \int_0^1 c^\star_{i,t} di$.

Let us make two remarks. First, the use of Walras's Law here makes the last requirement redundant: add consumers' budgets at each point in time and use the firm's problem, which has constant returns to scale, to see this.[11] Second, we formulate the firm's problem as static here; it is simpler. In it, note that whereas K and L are mere choice variables and need not have time subscripts, the solution does need to be dated.

5.4 Recursive equilibrium

Just like it is possible to study dynamic optimization problems with recursive methods—dynamic programming—it is also possible to define equilibria that way. We saw in the context of optimization that recursive methods make the object of study different: we look for functions, not sequences. For a maximization problem, having solved a dynamic program means that we have a function that we can apply to any value of its argument: the state variable (which can be a vector). A recursive equilibrium will also focus on functions.

[11] Walras's Law was applicable in the Arrow-Debreu formulation too, but then only once—not at each t.

To move slowly toward a general definition, let us first study how recursive methods can be used to define steady-state equilibria, i.e., equilibria where (aggregate) variables are constant over time. In the case of equilibria defined as sequences, we did not separately define steady-state equilibria: steady-state equilibria simply satisfy the definition of equilibria and have the additional property that all variables (at least aggregates) are constant over time. For the endowment economy, the case where all individuals' endowments are constant over time, all equilibria are steady-state equilibria. For the neoclassical economy, the economy's aggregate capital stock needs to have a certain starting value; otherwise, the equilibrium will not be in steady state from time zero. Steady-state equilibra defined using recursive methods are also just a subset of all equilibria. However, as we shall see, it is useful to define steady-state equilibria separately, as a first step.

5.4.1 Steady state

A steady-state equilibrium is one where the aggregate economy is at a rest point. This means, in particular, that prices are constant. Before proceeding, note that the present discussion also applies to the case where there is exact balanced growth, in which case the economy can be transformed into a stationary one whose steady state is then studied (in this case, wages will be growing along the balanced path but will be constant in the transformed economy).

Conceptually, now, for the definition of a steady-state equilibrium we will need to specify the constant prices and aggregates and to provide functions making clear that, at steady state, agents optimize: they consider behavior that is non-constant over time but choose to remain constant.

As before, we will go through definitions for different economies, beginning with a case without production.

The endowment economy

Consider, again, the economy with a continuum of consumers indexed by i, each now with and endowment that is constant over time, y_i. An equilibrium is defined as follows.

Definition 7 *A **recursive steady-state competitive equilibrium** is a q and a set of functions, $V_i^\star(a)$ and $g_i^\star(a)$, and asset values $a_i^\star$, for for each $i \in [0,1]$ such that*

1. *for each i, $V_i^\star(a)$ solves*

$$V_i(a) = \max_{a'} u(a + y_i - qa') + \beta V_i(a') \quad \forall a,$$

   with $g_i^\star(a)$ attaining the maximum on the right-hand side for all a;

2. *for each i, $g_i^\star(a_i^\star) = a_i^\star$; and*

3. *$\int_0^1 g_i^\star(a_i^\star)di = 0$.*

Here, the $V^\star$s and $g^\star$s capture individual optimization, given prices (a single q in this case). The second condition expresses stationarity of individual asset holdings; at the same time, it does put a sharp restriction on the function. We are, thus, allowing different individuals to start with—and maintain—different holdings of assets. Market clearing is captured in the third condition. Our equilibrium definition also specifies a distribution of asset holdings, which is required to be constant over time. A special case is that of a representative agent, where the second condition would read $g^\star(0) = 0$ and the third condition would be superfluous.

Equilibrium characterization The functional Euler equation, which we arrive at after taking the first-order condition in the dynamic program and then using the envelope theorem, reads, for all i,

$$qu'\left(a + y_i - qg_i(a)\right) = \beta u'\left(g_i(a) + y_i - qg_i(g_i(a))\right).$$

Using stationarity at $a = a_i$, we obtain

$$qu'\left(a_i(1-q) + y_i\right) = \beta u'\left(a_i(1-q) + y_i\right).$$

Hence, $q = \beta$. We can now solve for g_i from the functional Euler equation: it implies $a - qg_i(a) = g_i(a) - qg_i(g_i(a))$ and it is easy to see here that $g_i^\star(a) = a$ solves this functional equation for all i. No matter what asset level an agent has, they choose to keep it and just consume its accrued interest. This is the permanent-income result we have seen before, now in recursive form.

Given the simple form of the policy function, we immediately obtain the implied $V_i^\star$s. They satisfy $V_i^\star(a) = u(a(1-q) + y_i)/(1-\beta)$.

Having found all the equilibrium objects, we note, first, that the asset distribution can be chosen freely subject to market clearing and that all consumers have non-negative consumption. That is, this model has "no predictions" for long-run wealth distributions: any relative distribution of wealth is sustained over time. Since real-world wealth distribution have certain distinct features in common—regarding their overall shapes, across time and countries—this model therefore is not satisfactory. Richer models of wealth distribution that relate to data are therefore developed and discussed in Chapter 21. Second, we see that the shape of the utility function, u, does not play a role in any of the characterizations of the steady state. This is not true if there is balanced growth (in endowments); then, u needs to have the usual balanced-growth shape.

The neoclassical growth economy

Turning to the neoclassical growth model with optimal saving, let us proceed immediately to the definition of steady-state equilibrium, of course with the assumption that TFP, A_t, is constant over time.[12] We will again consider a continuum of consumers. For simplicity, we will assume that labor supply is exogenous and equal to e_i for agent i, with $\int_0^1 e_i di = 1$.

[12] Alternatively, it is growing at a constant rate, in which case the analysis requires a transformation of variables.

Definition 8 *A **recursive steady-state competitive equilibrium** consists of scalars r and w and a set of functions, $V_i^\star(k)$ and $g_i^\star(k)$, and capital holdings $k_i^\star$, for for each $i \in [0,1]$ such that*

1. *for each i, $V_i^\star(k)$ solves*

$$V_i(k) = \max_{k'} u((1-\delta+r)k + we_i - k') + \beta V_i(k') \quad \forall k,$$

   with $g_i^\star(k)$ attaining the maximum on the right-hand side for all k;

2. *$(K^\star, 1)$ solves $\max_{K,L} F(K,L) - rK - wL$, where $\int_0^1 k_i^\star di = K^\star$; and*

3. *for each i, $g_i^\star(k_i^\star) = k_i^\star$.*

We see a close similarity between the agent's problem here and that in the endowment economy: there exogenous labor income and a constant interest rate. There are two prices, r and w, but as we shall see they are closely related. Finally, we see that the market-clearing condition is different: assets sum up to the aggregate capital stock.

Equilibrium characterization For reasons identical to those used in the case of the endowment economy, we obtain $\beta = 1 - \delta + r$. This equation determines r. Then we know from firm maximization that $r = F_1(K^\star, 1)$, where $F_i(\cdot,\cdot)$ represents the partial derivative with respect to ith argument. This equation determines $K^\star$. We also know that $w = F_2(K^\star, 1)$, which gives us w. The distribution of capital is indeterminate, subject to adding up to $K^\star$. The policy and value functions are given by $g_i^\star(k) = k$ and $V_i^\star(k) = u(k(r-\delta) + we_i)/(1-\beta)$, respectively, for all k and i.

5.4.2 Dynamics

Turning to a full equilibrium, we now need to find a way to express, using functions, how prices and aggregates move over time. We will consider the neoclassical growth model only.[13] We will assume that TFP is constant, so as to isolate how these price and quantity movements are endogenous. We will, again, consider a continuum of households, but now assume as a benchmark that their asset holdings and labor productivities are the same. This means that the consumer we look at can be thought of as a representative agent from the outset, i.e., one among a [0,1] continuum of identical agents. We will revisit the question of actual heterogeneity at the end. We also begin with the assumption that labor supply is exogenous and equal to 1.

Recall that recursive methods involve expressing outcomes as functions of *state variables*. A state variable has to be both relevant and predetermined. So, in our economy, what determines prices and aggregates over time? It is instructive to consider period 0: what determines prices at that point in time? In the neoclassical model r_0 and w_0 are determined by the capital/labor ratio, through the marginal products of capital and labor. Hence, at the

[13] Endowment economies can be considered as well but do not involve the same core issues as the neoclassical model allows us to illustrate.

very least these two prices will depend on the period-0 capital stock in the economy: besides being relevant, the aggregate capital stock is also predetermined. Is there another variable that qualifies as a state, determining prices? No. Hence, we conclude that two functions $r(K)$ and $w(K)$ need to be part of our equilibrium.[14]

Agents will thus take as given $r(K)$ and $w(K)$ when making their decisions. But, given that prices will move over time, how do they know what prices will prevail in the future? For this they need the equivalent of the planner's decision rule for capital: in the context of a recursive equilibrium, we will label it a law of motion, $K' = G(K)$. Thus, agents take G, r, and w as given functions and then solve their dynamic programming problems. What will these problems look like? As in the case of the steady-state equilibrium we need to allow the agent's choice of capital to deviate from that in equilibrium. Therefore the dynamic program must read

$$V(k, K) = \max_{k'} u((1 - \delta + r(K))k + w(K) - k') + \beta V(k', G(K)) \quad \forall (k, K), \tag{5.23}$$

where k is an individual agent's capital and K is the aggregate capital stock, which they take as given.

We now define equilibrium:

Definition 9 *A **recursive competitive equilibrium** consists of functions $r(K)$, $w(K)$, $G^\star(K)$, $V^\star(k, K)$, and $g^\star(k, K)$ such that*

1. *$V^\star(k, K)$ solves (5.23), for $G = G^\star$ and $g^\star(k, K)$ attains the maximum in this problem;*
2. *for all K, $r(K) = F_1(K, 1)$ and $w(K) = F_2(K, 1)$; and*
3. *$G^\star(K) = g^\star(K, K)$ for all K.*

The second condition does not state profit maximization of the firm explicitly but just uses the first-order conditions.[15] The third condition, labeled *consistency*, works as a market clearing condition: it requires that the evolution of the aggregate capital stock is consistent with the choices of individuals when they each hold the same level of capital.

Is the resource constraint met in our definition? It is. The representative consumer will, in a recursive competitive equilibrium, consume $K(1 - \delta + r(K)) + w(K) - G^\star(K)$, which, from the firm's problem and F being CRS, equals $F(K, 1) - K' + K(1 - \delta)$, i.e., output minus investment.

Equilibrium characterization Let us also derive the functional Euler equation of the agent. Taking first-order conditions and applying the envelope theorem (for a marginal change in k), we obtain (dropping *s for convenience) that,for all (k, K),

$$\begin{aligned} &u'\left((1 - \delta + r(K))k + w(K) - g(k, K)\right) = \\ &\quad \beta u'\left(g(k, K)(1 - \delta + r(G(K))) + w(G(K)) - g(g(k, K), G(K))\right)\left[1 - \delta + r(G(K))\right]. \end{aligned}$$

[14] As an alternative to looking at time 0 to gain intuition, consider the planning problem of the economy under study: the state variable relevant to the planner will then be a state in the recursive equilibrium.

[15] Alternatively, state that $r(K)$ and $w(K)$ are such that $(K, 1)$ solves $\max_{k,\ell} F(k, \ell) - r(K)k - w(K)\ell$ for all K.

Given r, w, and G, this functional equation determines $g(k, K)$: the behavior of individual saving when the individual has k, possibly different from aggregate capital, K. Can g be solved for explicitly? In general, no. However, when u is a power function, again in line with our general balanced-growth requirements, it is possible to show that it takes the form $g(k, K) = \mu(K) + \lambda(K)k$.[16] Individual saving is *linear* in the own holdings of capital: the marginal propensity to save is $\lambda(K)$, i.e., independent of the level of k (it only depends on the aggregate capital stock). This also means that if we were to consider several consumers and distribute capital among them, how we distribute it would not matter for aggregate saving. In other words, we have *aggregation* if u is a power function. Thus, it is not restrictive to consider a representative agent. This result holds also if different agents have different endowments of labor: then $g_i(k, K) = \mu_i(K) + \lambda(K)k$, i.e., the intercept will vary across individuals but the slope will not. To conclude, if an economy with a power utility function has a nontrivial distribution of capital among agents, the aggregate law of motion will only depend on aggregate capital. That is, K is still the aggregate state variable: the distribution of capital, though predetermined, is not relevant for understanding how prices are determined.

Turning to how one solves for aggregates, we can also evaluate the above Euler equation at $k = K$ and use the fact that the resource constraint will hold, along with the expression for the rental rate function, to obtain

$$\begin{aligned} &u'\left(K(1-\delta)K + F(K,1) - G(K)\right) = \\ &\quad \beta u'\left(G(K)(1-\delta) + F(G(K),1) - G(G(K))\right)\left[1 - \delta + F_1(G(K),1)\right]. \end{aligned}$$

This functional equation solves for the evolution of the capital stock. We note that it coincides with the functional equation of the planner's problem for the same economy: equation (4.26) of our optimization chapter. As in that case, this equation has to be solved numerically unless u is logarithmic, F is Cobb-Douglas, and $\delta = 1$.

Before concluding, let us consider the economy with valued leisure: agents have utility functions $u(c) - v(\ell)$. Now aggregate labor supply is endogenous. It is not predetermined in a given period, so it is not a state variable (neither for the individual nor for the aggregate). Thus, our state is still (k, K). What is needed now, however, is a function $L = H(K)$ specifying how aggregate labor supply depends on the aggregate state. Similarly, on the individual level, we need $\ell = h(k, K)$ to denote the policy function for the choice of labor. The equilibrium, again for the representative-agent case, becomes

Definition 10 *A **recursive competitive equilibrium** for the economy with valued leisure consists of functions $r(K)$, $w(K)$, $G^\star(K)$, $H^\star(K)$, $V^\star(k, K)$, $g^\star(k, K)$, and $h^\star(k, K)$ such that*

1. *$V^\star(k, K)$ solves*

$$V(k, K) = \max_{k', \ell} u((1 - \delta + r(K))k + w(K)\ell - k') - v(\ell) + \beta V(k', G^\star(K)) \quad \forall (k, K).$$

and $k' = g^\star(k, K)$ and $\ell = h^\star(k, K)$ attain the maximum in this problem;

[16] See Appendix 5.A for a proof and a more extensive discussion of conditions under which the solution has this form.

2. *for all* K, $r(K) = F_1(K, H^\star(K))$ *and* $w(K) = F_2(K, H^\star(K))$; *and*

3. $G^\star(K) = g^\star(K, K)$ *and* $H^\star(K) = h^\star(K, K)$ *for all* K.

This is a straightforward extension of the definition without valued leisure. Notice that the agent does not use $H^\star$ in the maximization problem; $G^\star$ suffices, because r and w now capture how the labor input changes with K.

It is straightforward, but somewhat tedious, to derive the functional Euler equations; there will now be an intertemporal condition too determining labor supply. Again, evaluated at $k = K$, one finds that the conditions are identical to those of the corresponding planner's problem.

Will this economy deliver aggregation too, once u is of the power form? As already alluded to above in Section 5.2.3, the answer is no, unless labor productivities differ across agents too and the ratio of the initial capital holding to labor productivity is the same across all agents.[17] If not, the aggregate state variable necessarily becomes the vector of capital holdings of all agents, not just the sum of these holdings.

5.5 Overlapping generations

We now consider overlapping generations models. The defining feature of these models is that agents live for a finite length of time but the economy continues after their death with new generations of agents. Historically, overlapping-generations (OG) models have played an important role in macroeconomics. They were first introduced by Allais (1947) and later used for a large variety of purposes: for understanding why fiat money has value and the potential need for a social security system (Samuelson, 1958a) and for understanding government debt (see Diamond, 1965). Interestingly, although most of his later work used dynastic settings, Lucas's path-breaking (see Lucas, 1972) paper on the Phillips curve is using an OG model. We will revisit these applications later in the text. In this chapter, we will merely focus on some features that make OG market economies different than those studied above.

We will restrict attention to the simplest version of an OG setting: one where, each period, a cohort of people are born and then live for two periods only. That way, cohorts overlap, but only for one period. This is the so-called the two-period life OG setting. Clearly, the two-period life case is limited in its applicability: for example, it cannot be used for quantitative studies of the business cycle, as business cycles occur rather frequently (a time period in the OG model should perhaps be thought of as 25 or 30 years). However, it is of course possible to construct OG models where people live for an arbitrary (finite) number of periods and many of the special properties of such models are inherited from the two-period life case.

In our benchmark, we will assume—as is the case in the most commonly used OG model—that people, though they have children, do not give bequests or any other gifts to them. Relatedly, they express no altruism toward their offspring and thus maximize their own utility only, defined over consumption (and possibly leisure) in their two periods of life. The maximization problems are, then, conceptually simple; the finite-horizon settings studied in

[17] In this case, this ratio will also stay the same over time.

Section 4.2, in particular that in 4.2.1, can be immediately applied. The maximization problems, moreover, are often called *life-cycle models*, emphasizing that individuals go through different phases in life and, in particular, that young and old individuals have different time horizons. Here, we will limit heterogeneity to age and not consider further differences between people. Thus within each cohort, all agents are identical.

Defining equilibria for OG models is straightforward, too. As for the dynastic setting, there are three methods—sequence-based (AD and sequential-trade) and recursive definitions of equilibria—but here, for convenience only, we will focus exclusively on the sequential-trade setting. We will see that not only are the maximization problems simpler in OG models but computing equilibria is more straightforward too. However, what we will find is that equilibria for OG models can (but do not necessarily) have peculiar features. For example, equilibria may not be Pareto optimal. In addition, there may be more than one equilibrium. In this chapter, we will mainly set things up; welfare properties will then be studied carefully in the next chapter and applications will be discussed in later chapters.

As before, we begin with the endowment case. We will then turn to the neoclassical growth model. Lastly, we will discuss introducing altruism and bequests. Throughout, we abstract from population growth.

5.5.1 The endowment economy

Let us assume that the cohort born at t has utility given by

$$u_t(c_y,\, c_o) = u(c_y) + \beta u(c_o),$$

where we evaluate at two arbitrary consumption levels (c_y, c_o) when young and old, respectively. The preferences of generation $t = -1$, who are old as time begins, are similarly represented by $u_{-1}(c) = u(c)$.

We consider endowment sequences given by $(\omega_{y,t}, \omega_{o,t+1})$ for cohort t: for all t, where t is the time period, $\omega_{y,t}$ is the endowment of the young and $\omega_{o,t}$ the endowment of the old (who are of cohort $t-1$) in that time period.

A sequential equilibrium is defined in much the same way as in our endowment economy with dynastic agents.

Definition 11 *A **sequential competitive equilibrium** is a set of sequences $\left\{c^\star_{i,t}\right\}_{t=0}^{\infty}$, for each $i \in \{y, o\}$, $\left\{a^\star_{t+1}\right\}_{t=0}^{\infty}$, and $\{q_t\}_{t=0}^{\infty}$ such that*

1. *for each $t > 0$, $\left(c^\star_{y,t}, c^\star_{o,t+1}, a^\star_{t+1}\right)$ solves*

$$\max_{c_y, c_o, a'} u(c_y) + \beta u(c_o) \quad \text{subject to} \quad c_y + q_t a' = \omega_{y,t} \text{ and } c_o = \omega_{o,t+1} + a'$$

   and $c_{o,0} = \omega_{o,0}$.

2. *for all $t \geq 0$, $c^\star_{y,t} + c^\star_{o,t} = \omega_{y,t} + \omega_{o,t}$.*

The last requirement—goods market clearing—can equivalently, using consumers' budgets, be written $a^\star_{t+1} = 0$ for all t. To see this, begin in period 0 and roll forward: resource

feasibility at time 0 means, when summing cohort 0's budget and cohort 1's first-period budget, that $a_1^\star$ must equal 0. Then the same procedure for next period delivers $a_2^\star = 0$, and so on. Intuitively, this is obvious: any given cohort fundamentally only have endowment income, and saving (or borrowing) when young must mean that another cohort is on the other side of that transaction; but it cannot be the old who is now alive during their last period, and it cannot be the young of next cohort, since they have not been born yet.

Equilibrium characterization Having established $a_{t+1}^\star = 0$, a result that is true also in the dynastic model if all agents have the same endowments, we conclude that $(c_{y,t}^\star, c_{o,t}^\star) = (\omega_{y,t}, \omega_{o,t})$: autarky. These results are in line with what we found in the dynastic model. Also, as before, solving for the price only involves evaluating the consumer's first-order (Euler) condition, which—evaluated at autarky—reads

$$q_t u'(\omega_{y,t}) = \beta u'(\omega_{o,t+1})$$

at time t. If we impose $u(c) = (c^{1-\sigma} - 1)/(1-\sigma)$, so that (in a slightly more general model) we obtain results consistent with balanced growth, we conclude that

$$q_t = \beta \left(\frac{\omega_{y,t}}{\omega_{o,t+1}} \right)^\sigma . \tag{5.24}$$

If the endowments are stationary, so that $\omega_{y,t}$ and $\omega_{o,t}$ do not depend on time, we obtain that q_t is constant and equal to $\beta(\omega_y/\omega_o)^\sigma$. Thus, if the endowment when young does not equal the endowment when old, the interest rate will not just reflect the utility discount factor β but also the shape of *life-cycle income*. In particular, if $\omega_y > \omega_o$, which makes sense if we identify y with "working" and o with "being retired," then we can even obtain $q > 1$, i.e., negative real interest rates, in this model. This is not possible to obtain in the dynastic endowment model (unless aggregate endowments fall over time at a constant, and high enough, rate).

The intuition for the possibility of negative interest rates in the OG model is that when life-cycle endowments decline over time (at a high enough rate), consumption is marginally more valuable in the future than now in the absence of being able to smooth income over time. And in the two-period life OG model no such smoothing is feasible. In an OG model where people live for three periods there could be borrowing/lending between the young and the middle-aged. It would be reasonable to think that the life-cycle endowment pattern then has $\omega_y < \omega_m$ and $\omega_m > \omega_o$, where ω_m is the endowment of the middle-aged. The young then could borrow from the currently middle-aged, who want to lend. Thus, some smoothing would be obtained and the discount factor would play a more prominent role again; but full smoothing would not necessarily materialize and, thus, the OG model continues to give different predictions than does the dynastic model. Overall endowment growth would also affect the result in the direction of producing higher real interest rates.

Finally, note that it would be beneficial to transfer resources from the young to the old, in the case where $\omega_y > \omega_o$. If *all* young transfer to the current old then it is even possible for *all* generations to benefit. This kind of transfer can be thought of as a government-run pay-as-you-go pension scheme; in fact, this is a key early use of the OG model. Thus, we have an indication that the market is not efficient here. We will discuss this issue at much greater length in our welfare chapter: Chapter 6.

5.5.2 The neoclassical growth economy

In the two-period life OG model, the introduction of capital allows consumers to actually life-cycle save. They are endowed with labor income when young and labor income when old and, as before, have no initial assets. So if their income as young is higher than their income when old, they would buy capital when young and rent it out, as old agents, to firms. At time 0, the capital stock is thus owned by the old at that time.

For concreteness, we let the labor productivity (or alternatively time endowment) of the young and old at all times be e_y and e_o, respectively, with $e_y + e_0 = 1$. We also restrict attention to a stationary production function. Growth in productivity is straightforward to introduce and it is not essential for the key points here. We have the following (where we use a representative agent within each cohort to economize on notation).

Definition 12 *A **sequential competitive equilibrium** is a set of sequences $\left\{c^\star_{i,t}\right\}_{t=0}^\infty$, for each $i \in \{y, o\}$, $\left\{k^\star_{t+1}\right\}_{t=0}^\infty$, $\{r_t\}_{t=0}^\infty$, and $\{w_t\}_{t=0}^\infty$ such that*

1. *for each $t > 0$, $\left(c^\star_{y,t}, c^\star_{o,t+1}, k^\star_{t+1}\right)$ solves*

   $$\max_{c_y, c_o, k'} u(c_y) + \beta u(c_o) \quad \text{subject to} \quad c_y + k' = w_t e_y \text{ and } c_o = w_{t+1} e_o + (1 - \delta + r_{t+1})k'$$

   and $c_{o,0} = w_0 e_o + (1 - \delta + r_0)k_0$;

2. *for all t, $r_t = F_1(k^\star_t, 1)$ and $w_t = F_2(k^\star_t, 1)$, with $k^\star_0 \equiv k_0$; and*

3. *for all $t \geq 0$, $c^\star_{y,t} + c^\star_{o,t} + k^\star_{t+1} = F(k^\star_t, 1) + (1 - \delta)k^\star_t$.*

This definition is in line with the one for a dynastic economy with the one difference that there are two types of agents, only one of which saves in capital at any point in time.

Equilibrium characterization The focus is on the individual problem. The Euler equation is entirely standard-looking, so let us proceed immediately to evaluate it after expressing prices as a function of capital stocks:

$$\begin{aligned} &u'\left(e_y F_2(k^\star_t, 1) - k^\star_{t+1}\right) = \\ &\quad \beta u'\left(e_o F_2(k^\star_{t+1}, 1) + (1 - \delta + F_1(k^\star_{t+1}, 1))k^\star_{t+1}\right)\left(1 - \delta + F_1(k^\star_{t+1}, 1)\right). \end{aligned}$$

Conditional on a value for $k^\star_t$, this equation solves for $k^\star_{t+1}$. Therefore, we have a very different dynamic system than in the dynastic model, where we always obtained a second-order difference equation. That is, here, we can "solve forward": start with k_0, solve for $k^\star_1$, and so on.

Second, what does a steady state look like? Letting $\bar{k}$ denote a steady state, we obtain

$$u'\left(e_y F_2(\bar{k}, 1) - \bar{k}\right) = \beta u'\left(e_o F_2(\bar{k}, 1) + (1 - \delta + F_1(\bar{k}, 1))\bar{k}\right)\left(1 - \delta + F_1(\bar{k}, 1)\right).$$

That is, we do not, in general, obtain $\beta\left(1 - \delta + F_1(\bar{k}, 1)\right) = 1$, since consumption may not end up being fully smoothed. Again, thus, we see that OG models have qualitatively different implications for long-run interest rates: they can, depending on parameter values (intuitively,

depending on the demand for saving given the life-cycle structure, and depending on firm's demand for capital) be either higher or lower than $1/\beta$ (in gross terms). We invite the reader to fully solve the model where u is logarithmic, $e_y = 1$ and $e_o = 0$, F is Cobb-Douglas, and $\delta = 1$ and verify that (i) capital's dynamics will be log-linear and converge monotonically to a steady state; and (ii) the gross steady-state interest rate will be $\alpha(1+\beta)/(1-\alpha)$, a number which is less than one if α is low enough.

Finally, we also note that although the solution for the model's dynamics involves only a first-order difference equation, it is not immediately obvious that, for each k_t, there is a unique value k_{t+1} solving the Euler equation. In our dynastic model, we obtain a unique equilibrium so long as it can be obtained as a solution to a planner's problem, which we know to be unique under standard assumptions. Here, there is no immediate connection to a planner's problem.[18]

5.5.3 Some model comparisons

We have studied the OG model in its two-period life version. When consumers live for more than two periods, the model in some ways looks more like a dynastic model; as the time horizon gets longer, the life-cycle patterns can become less pronounced. Also, solving for equilibria in OG models with more than two-period lives involves difference equations that are of higher order than one. Still, however, they are conceptually different. The determination of the long-run real interest rate, for example, remains more complex than in the dynastic model, where very few parameters matter: without aggregate growth, the gross real interest rate is $1/\beta$, and with net growth at rate γ, it is $(1+\gamma)^\sigma/\beta$.

Random deaths There are other variants of the OG model. One is the *perpetual-youth* (or, perhaps, *sudden-death*) model. There, all individuals face a constant probability of death between t and $t+1$. Thus, a "lucky" individual can live very long, even forever. As a consequence, anybody alive at t has the same expected remaining lifetime. An individual who dies will be replaced by a new-born individual, so as to avoid a shrinking population, but again not as part of a dynasty: this model shares with the basic OG setting that no individual cares about children. When an individual dies with unspent assets—which is typical—then these assets can be either seized by the government or given as "random bequests" to the surviving population. A third alternative—all have been used in the literature—is that agents can write a form of annuity contracts whereby they obtain a higher than the (safe) market interest rate if they survive, in return for losing all the money at death; a "bank" on the other side of this transaction would then on average, if the contract is written with many individuals, obtain a safe return. The perpetual-youth model inherits some of the characteristics of the OG model but, at the same time, is a move toward a dynasty model: as the probability of survival approaches one, the model's features approach those of the pure dynasty setting.

[18] Note also that a planner's objective in the OG setting would need to involve a social welfare function across cohorts.

Warm glow One can also append an OG model with *bequest functions*. The idea here is that people do not exhibit altruism but, rather, they care about the act of giving money away (typically, to their children). This setting is referred to as one of *warm glow*: giving makes one happy (glow). To illustrate, consider the two-period life OG model. The preferences are now

$$u(c_y) + \beta u(c_o) + \varphi(b'),$$

where φ (like u) is an increasing, strictly concave function and b' is the amount of bequests given. Thus, the budget constraints for cohort t (assuming a stationary endowment economy for simplicity) read

$$c_{y,t} + q_t a_{t+1} = \omega_y + b_t \text{ and } c_{o,t+1} + b_{t+1} = \omega_o + a_{t+1}.$$

Here, b_t is the bequest inherited from cohort $t-1$ (the parents) and b_{t+1} is the amount bequeathed to cohort $t+1$. Market clearing could again be expressed by requiring that total consumption equal total endowments period by period; equivalently, one could simply require a_{t+1} to be zero at all times. Now notice that the old at time 0 have a non-trivial decision: maximize, by choice of $(c_{o,0}, b_0)$, $\beta u(c_{o,0}) + \varphi(b_0)$ subject to $c_{o,0} + b_0 = \omega_o$.

An alternative interpretation of this function is that the consumer derives utility from their end wealth: "dying with a bigger bank account balance" gives you higher utility. Regardless of interpretation, this kind of model can be seen as a *behavioral* model: agents no longer have preferences over consumption goods (and leisure) but are endowed with utility functions defined by the act of giving. A conceptually different approach is to assume that parents value their offspring's consumption paths; if they do, but they do so in a way that is inconsistent with the way the children value their own consumption paths, such preferences are said to capture impure altruism. A model with pure altruism is instead one where the parent cares about the child *the way the child cares about themselves*, in which case one can no longer label the model behavioral since now preferences are defined over goods. If this is true for all cohorts, we can specify the utility of cohort t as follows:

$$u(c_{y,t}) + \beta u(c_{o,t+1}) + \tilde{\beta}\varphi(b_{t+1}),$$

where

$$\varphi(b_{t+1}) = u(c_{y,t+1}) + \beta u(c_{o,t+2}) + \tilde{\beta}\varphi(b_{t+2}),$$

and so on. Notice here that $\tilde{\beta}$, the weight on the child's indirect utility, does not have to equal β. This reveals a recursive structure, which is most easily described with dynamic programming notation. Assuming a constant interest rate $1/q$ for simplicity, we obtain

$$\varphi(b) = \max_{c_y, c_o, a', b'} u(c_y) + \beta u(c_o) + \tilde{\beta}\varphi(b'),$$

subject to $c_y + qa' = \omega_y + b$ and $c_o + b' = \omega_o + a'$.

In sum, we have seen the perpetual-youth model, which looks more and more like the dynastic model as the survival probability goes to one. The warm glow model, on the other hand, can be seen as a behavioral OG model, except in the very special case where φ is (a constant times) the indirect utility function of the next cohort. Then, the model becomes

a *dynastic life-cycle model*: one where there are life-cycle features but where parents care about children in a (purely) altruistic way. In this very special case, thus, the basic long-run features are exactly those of the simpler dynasty model we use in our benchmark. The long-run interest rate is now given by $1/\tilde{\beta}$, and so on.

Chapter 6

Welfare

The preceding chapter presented examples where the competitive equilibrium allocation is the same as the solution to a social planner's problem. These examples illustrate the First Welfare Theorem of economics, which gives conditions under which a competitive equilibrium is Pareto optimal. The First Welfare Theorem is a remarkable result—arguably the most impressive insight that economics has come up with so far. Market mechanisms can effectively coordinate the activities of large numbers of consumers and firms, who may be spread across the world and alive at different times. Amazingly, under the conditions of the First Welfare Theorem (FWT), the actions of these disparate and heterogeneous people are coordinated in a way that is socially optimal even though each individual is simply pursuing their private interests. This is Adam Smith's invisible hand.[1]

Under the FWT, a competitive equilibrium is Pareto optimal. We focus on the concept of Pareto optimality because it is widely used as a "minimal" welfare criterion. It is, in particular, silent on the distributional implications of an allocation. For example, an allocation with extreme inequality can nevertheless be Pareto optimal. At times, macroeconomists assume more specific social welfare functions that express a preference for more equal distributions of resources and we will comment upon this briefly.

While markets can work very well, they can also fail, which is to say that the competitive equilibrium may not be Pareto optimal. Such failures of the FWT can arise for a number of reasons, which we will discuss in detail. Our aim is to understand when market economies work well and when they do not. Indeed, much of modern macroeconomics is concerned with understanding the potential market failures that affect the economy and understanding the public policies that might lead to better outcomes. An understanding of the nature of market failure and how it can be corrected is an important part of evaluating the potential benefits of policy proposals.

After first reviewing the relationship between Pareto optimality and competitive equilibrium (the FWT) in an abstract, general setting, we map it into the more applied macroeconomic settings used in our text. We then go through the most common market failures that arise in macroeconomic analysis and indicate how each of these failures involve departures from the assumptions underlying the welfare theorem. We end the chapter with a short discussion of the role of government policy.

[1] An inspiring description of the power of markets can be found in the short video available at `https://www.youtube.com/watch?v=67tHtpac5ws`

6.1 The First Welfare Theorem

The First Welfare Theorem says that a competitive equilibrium is Pareto optimal. To make this statement more precise and to give a sketch of the proof we will consider an abstract economy with many different goods. These goods could be different commodities or goods at different dates. Later we will describe how different specific economic environments relate to this abstract economy.

There is a set $\mathcal{I}$ of different consumers, each indexed by $i \in \mathcal{I}$, and a set $\mathcal{J}$ of firms, each indexed by $j \in \mathcal{J}$. We think of $\mathcal{I}$ and $\mathcal{J}$ as finite but we will consider them to be infinite in some extensions that we discuss below. Let x be a vector of consumption levels of different goods that are traded. The length of the vector thus specifies how many markets there are in the economy. For now, we will think of the vector length as finite, but it will be relevant to consider the possibility of infinite vectors later. Moreover, the vector is, at least for now, allowed to have negative elements. Similarly, y is a vector of production levels of the same goods and ω is a vector of exogenous endowment levels of the goods. Consumer i is endowed with ω_i and consumes x_i while firm j produces y_j. Thus, our resource constraint reads

$$\sum_{i \in \mathcal{I}} x_i \leq \sum_{j \in \mathcal{J}} y_j + \sum_{i \in \mathcal{I}} \omega_i. \tag{6.1}$$

This inequality applies to each element of the vector so, for every good, the total amount consumed cannot exceed the endowment plus the amount produced by all the firms.

Let X_i be a set of vectors that are feasible for agent i to consume. For example, if there are two commodities, then $X_i = \mathbb{R}^2_+$ would rule out consuming a negative amount of either good. Firms have production possibility sets denoted by Y_j. We assume that consumers maximize utility taking their budget constraint, prices, and consumption possibility sets as given. Utility maximization means they choose a consumption bundle that is not preference-dominated (as defined by a preference ordering $\succeq_i$) by any other choice available to them. Similarly, firms maximize profits, taking prices as given, by choosing a production plan in their production possibility set Y_j. Firm profits are given to the consumers that own the firms with $\theta_{i,j}$ denoting consumer i's share of firm j. Each firm is wholly owned by the consumers so $\sum_i \theta_{ij} = 1$. Finally, let p be the vector of prices for each good.

Key assumption We will use the weakest assumption under which the First Welfare Theorem holds: local non-satiation (LNS). LNS expresses that each consumer can always be made better off by an infinitesimally higher consumption of some good.[2]

A competitive equilibrium A competitive equilibrium is a consumption allocation $\{x_i^\star\}_{\forall i}$, a production allocation $\{y_j^\star\}_{\forall j}$, and a price system $p^\star$ such that

1. for each $i \in \mathcal{I}$, the consumption choice $x_i^\star$ is in X_i and there is no $x \in X_i$ such that $x \succ_i x_i^\star$ and $px \leq px_i^\star = p\omega_i + \sum_j \theta_{i,j} p y_j$;
2. for each $j \in \mathcal{J}$, $y_j^\star \in Y_j$ and there is no $y \in Y_j$ such that $py > py_j^\star$;

[2]This notion also presumes that the consumption possibility sets X_i allow small movements in at least one desirable direction.

3. and the market for each good clears (equation (6.1) holds with equality).

Theorem 6.1 (The First Welfare Theorem) *An allocation that is part of a competitive equilibrium is Pareto optimal.*

Proof. The proof is by contradiction. So suppose there exists an allocation $\{\tilde{x}_i\}_{\forall i}$, $\{\tilde{y}_j\}_{\forall j}$ that is feasible (these values are all in their possibility sets and the resource constraint is satisfied) and that Pareto dominates the allocation of the given equilibrium. Then, by definition,

$$\forall i \in \mathcal{I} : \tilde{x}_i \succeq_i x_i^\star,$$
$$\exists \tilde{i} \in \mathcal{I} : \tilde{x}_{\tilde{i}} \succ_{\tilde{i}} x_{\tilde{i}}^\star.$$

Now the first property of a competitive equilibrium combined with LNS can be used to conclude that

$$p\tilde{x}_i \geq px_i^\star$$

holds for all i. If it were strictly cheaper to buy the new allocation, the consumer could have spent more to improve on the existing choice and hence that choice could not have been optimal for the consumer. Moreover, for $\tilde{i}$ it must be that

$$p\tilde{x}_{\tilde{i}} > px_{\tilde{i}}^\star,$$

again because otherwise the original allocation must not have involved consumer optimization.

Summing the budget constraints of all consumers we have

$$p\sum_{i\in\mathcal{I}} x_i^\star = p\omega + p\sum_{j\in\mathcal{J}} y_j^\star,$$

where $\omega \equiv \sum_i \omega_i$ and we have used the fact that $\sum_i \theta_{ij} = 1\ \forall j$. By the arguments above, the alternative consumption allocation is more expensive

$$p\sum_{i\in\mathcal{I}} \tilde{x}_i > p\omega + p\sum_{j\in\mathcal{J}} y_j^\star.$$

Furthermore since $y_j^\star$ is profit maximizing for each j, it must be that $\sum_j p\tilde{y}_j \leq \sum_j py_j^\star$ so we have

$$p\sum_{i\in\mathcal{I}} \tilde{x}_i > p\omega + p\sum_{j\in\mathcal{J}} \tilde{y}_j.$$

In addition, the alternative allocation is resource feasible. Multiplying equation (6.1) by p we obtain

$$p\sum_{i\in\mathcal{I}} \tilde{x}_i \leq p\omega + p\sum_{j\in\mathcal{J}} \tilde{y}_j.$$

These last two inequalities contradict one another. ■

Mapping the setting into our macroeconomic models The power of the abstract proof above is that it can be applied to a large number of contexts. Let us first consider a static macroeconomic model with one agent with endowments of capital and labor equal to k and 1, and a neoclassical, constant-returns to scale production function that produces a consumption good. Preferences could be represented by a standard utility function, which satisfies LNS because it is strictly increasing. Then $x \in \mathbb{R}^3_+$ and a typical consumer choice would be $(c, 0, 0)$. Firms would have $(y, -k, -l) \in \mathbb{R}_+ \times \mathbb{R}^2_-$, the endowment vector would be $\omega = (0, k, 1)$, and the normalized price vector would be $(1, r, w)$.

The argument can easily be extended to include different types of goods. For example, if consumers value leisure we can simply treat them as "buying leisure" so that is just another good for them to consume. Other intermediate goods can be included too, as well as other resources, such as land. The theorem also applies to endowment economies by specifying the production possibility sets to only include the zero vector.

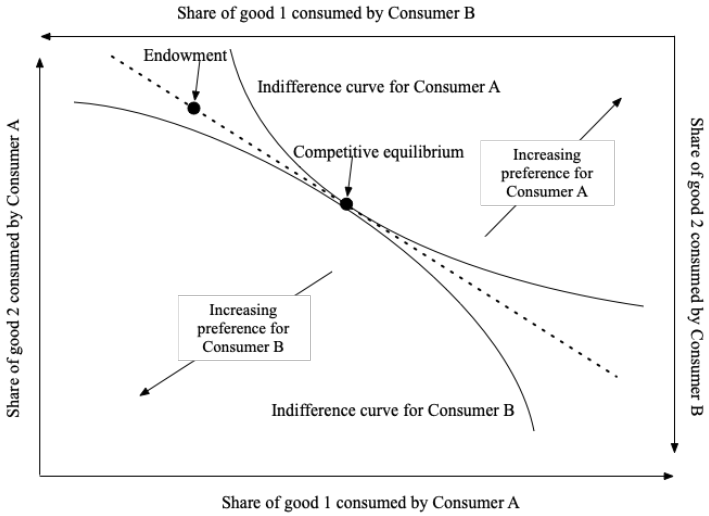
In a dynamic setting with $T < \infty$ time periods, goods and services at different dates are simply additional market goods and all the vectors become correspondingly longer. One can, in particular, define the vectors as simply containing T times the number of elements in the static model. As we will see below, the argument also applies to infinite horizon models but with some additional caveats. In the next chapter we will discuss models with uncertainty where goods are indexed not just by time, but also by the state of the world. The First Welfare Theorem applies to those settings as well.

Infinite horizon models We start by considering a case with a finite number of infinitely-lived consumers. Later in the chapter we will discuss an overlapping-generations economy where time is infinite but each consumer has a finite life.

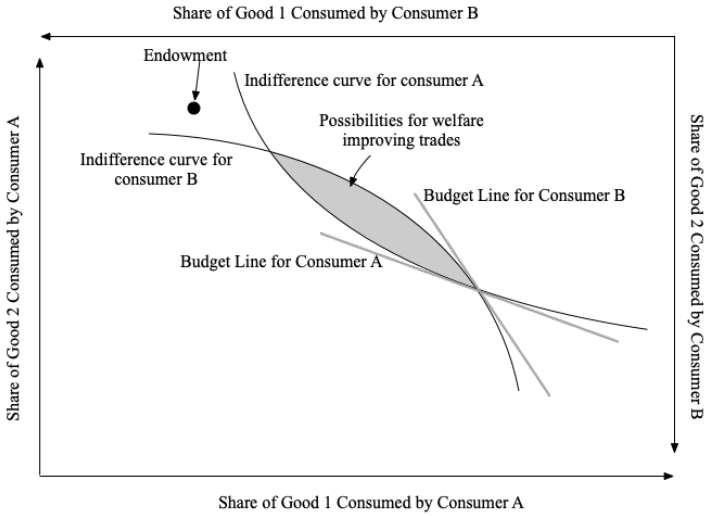
All the vectors in the proof above are infinite-dimensional but the competitive equilibrium is defined as before. A key point here is that for an equilibrium to exist—especially for it to satisfy its first property—it would have to be that the value of total expenditures of the consumer is finite, i.e., the dot product of the infinite price vector and the infinite quantity vector is finite. For an infinite sum, this would require, unless quantities go to zero, that prices fall sufficiently fast. Typically in our models, prices fall asymptotically at a geometric rate (equal to $\beta < 1$ if there is no growth), which for constant quantities imply a finite total value. As we shall see below, although this property holds in dynastic models, it does not hold in some overlapping-generations models. This means that overlapping-generation models can have very different welfare properties.

Intuition One issue underlying the First Welfare Theorem is the fact that all consumers and firms face the same prices. For all consumers the ratio of the marginal utilities of any two goods will equal the ratio of prices and, as the prices are the same, they all have the same ratio of marginal utilities. It is therefore impossible to make a marginal change in the consumption allocation that results in a Pareto improvement. This point can be illustrated with an Edgeworth box as shown in the top panel of Figure 6.1.

The figure depicts an endowment economy with two consumers and two goods. Their endowments are indicated by the dot. Given relative prices, we can draw a budget line as shown by the dashed line. Notice that each consumer will choose a consumption bundle



(a) Efficient equilibrium.



(b) Distorted market.

Figure 6.1: Edgeworth boxes of an endowment economy with two goods and two consumers.

Note: Panel (a) shows an efficient equilibrium in which both consumers face the same prices and have the same budget line. Panel (b) shows a distorted market where the two consumers face different prices.

that makes their indifference curve tangent to the budget line. When we impose market clearing, the points representing their consumption bundles in the Edgeworth box must coincide because what is not consumed by Consumer A must be consumed by Consumer B. Such a situation is marked as a competitive equilibrium in the figure. Notice that the indifference curves are tangent to the budget line and therefore also to each other. Because the indifference curves do not intersect, there is no way to move Consumer A to a higher indifference curve without harming Consumer B. Now consider the lower panel of the figure where we have imagined the two consumers face different prices. We will discuss several reasons this could occur, but a simple one is that Consumer A faces a tax on consuming one of the goods while Consumer B does not. If the consumers face different after-tax prices they will have different budget lines. As before each consumer chooses a consumption bundle where their indifference curve is tangent to their budget line. As the budget lines are different, the indifference curves now intersect and there are alternative allocations that yield higher welfare as shown by the shaded area in the figure.

Similar logic applies to firm profit maximization. Profit maximization leads firms to set the marginal rate of transformation equal to the ratio of prices. As firms face the same prices as consumers, the marginal rate of transformation between any two goods will therefore be equal to the marginal rate of substitution between any two goods. A marginal change in the production allocation therefore cannot transform goods in a way that improves consumer welfare. This point can be visualized as point of tangency between the indifference curves of the consumers (remember they all have the same slope in equilibrium) and the production possibility frontier. A marginal change along the production possibility frontier does not increase consumer welfare.

These intuitive arguments are not as powerful as the more abstract proof above because they apply to marginal changes in allocations while the abstract proof is global. Nevertheless, they can be helpful in developing an intuitive understanding for when the First Welfare Theorem will hold.

How do these ideas apply to macroeconomic models? In the dynastic growth model with optimal saving, the Euler equation sets the marginal rate of substitution between goods at t and $t+1$ equal to the relative price between them, i.e., the real interest rate. As all consumers face the same interest rate, they all have a common marginal rate of substitution between t and $t+1$ goods. Moreover, as firms rent capital from households at a rental rate that is equal to the real interest rate, the marginal rate of transformation between goods at t and $t+1$ is equal to the households' marginal rates of substitution. Similarly, if we consider the model with elastic labor supply, the firm will produce at a point where the marginal product of labor is equal to the wage. The consumer will supply labor such that the marginal rate of substitution between consumption and leisure is equal to the wage. Eliminating the wage from these two optimality conditions gives us that the marginal rate of substitution between goods and leisure is equal to the marginal rate of transformation between goods and leisure.

6.2 Tracing out the Pareto frontier

One useful way to construct a Pareto optimal allocation is to solve a social planner's problem in which the planner maximizes the weighted sum of the agents' utilities. A solution to such

a problem must be Pareto optimal because if it were not, the Pareto dominating allocation would increase the planner's objective function. As we will explain, the solution to this planner's problem is closely related to the concept of competitive equilibrium.

For the sake of simplicity, assume a set of infinitely-lived consumers, indexed by $i \in \mathcal{I}$, live in an exchange economy. Each period, there is a single consumption good and each consumer receives an endowment $\omega_{i,t}$. Each consumer has preferences given by

$$U_i \equiv \sum_{t=0}^{\infty} \beta^t u(c_{i,t})$$

with $u'(c) > 0$ and $u''(c) < 0$. Let p_t be the date-0 price of the date-t good. A competitive equilibrium is a consumption allocation $\{c_{i,t}\}_{\forall i,t}$ and a price system $\{p_t\}_{\forall t}$ such that all consumers are maximizing utility taking prices as given and markets clear. Market clearing requires that $\sum_i c_{i,t} = \sum_i \omega_{i,t}$ for all t.

In the competitive equilibrium with date-0 trading, the consumers maximize their utility subject to the date-0 budget constraint. The Lagrangian of this problem is

$$\mathcal{L} = \sum_{t=0}^{\infty} \beta^t u(c_{i,t}) + \lambda_i \sum_{t=0}^{\infty} p_t(\omega_{i,t} - c_{i,t})$$

and the first-order condition for $c_{i,t}$ is

$$\beta^t u'(c_{i,t}) = \lambda_i p_t.$$

As the utility function is strictly concave, we can invert $u'(\cdot)$

$$c_{i,t} = (u')^{-1}\left(\beta^{-t}\lambda_i p_t\right), \tag{6.2}$$

which shows that knowledge of $\{\lambda_i\}_{\forall i}$ and $\{p_t\}_{\forall t}$ is enough to determine the entire consumption allocation. The Lagrange multiplier λ_i captures the shadow value of date-0 wealth and will be decreasing in date-0 wealth.

Now consider a social planner that seeks to maximize a weighted sum of the utilities of the consumers. The planner's objective is

$$\sum_{i \in \mathcal{I}} \mu_i U_i,$$

where μ_i is the weight on consumer i's utility. These weights are called Negishi weights or Pareto weights. The constraint on the planner is the resource constraint

$$\sum_{i \in \mathcal{I}} c_{i,t} \leq \sum_{i \in \mathcal{I}} \omega_{i,t},$$

which must hold at each date. The Lagrangian of this problem is

$$\mathcal{L} = \sum_{i \in \mathcal{I}} \mu_i \sum_{t=0}^{\infty} \beta^t u(c_{i,t}) + \sum_{t=0}^{\infty} \psi_t \sum_{i \in \mathcal{I}} (\omega_{i,t} - c_{i,t}),$$

where ψ_t is the Lagrange multiplier on the resource constraint at date t. The first-order condition of this problem with respect to $c_{i,t}$ is

$$\mu_i \beta^t u'(c_{i,t}) = \psi_t.$$

Inverting $u'(\cdot)$ as above yields

$$c_{i,t} = (u')^{-1}\left(\beta^{-t}\frac{1}{\mu_i}\psi_t\right), \tag{6.3}$$

which shows that knowledge of $\{\mu_i\}_{\forall i}$ and $\{\psi_t\}_{\forall t}$ is enough to determine the entire consumption allocation.

There is a clear symmetry between equations (6.2) and (6.3). If $\mu_i = 1/\lambda_i$ and $\psi_t = p_t$, the two equations will give rise to the same consumption allocation. Let's assume that $\mu_i = 1/\lambda_i$. It must then be the case that $\psi_t = p_t$ because in the competitive equilibrium markets must clear and in the solution to the planner's problem the aggregate resource constraint must hold, which means in both cases the total consumption must equal the total endowment. By using equations (6.2) and (6.3) we can see that the only way we acheive the same total consumption at date t is if $\psi_t = p_t$. It follows that if we have the right Negishi weights, the competitive equilibrium is optimal in the eyes of a planner that attaches more weight to those with higher date-0 wealth. This argument is closely related to the First Welfare Theorem: there are some Negishi weights that make the competitive equilibrium optimal for the planner, which implies the competitive equilibrium is Pareto optimal.

Now let's flip the argument in reverse. As we vary the Negishi weights, we will arrive at different consumption allocations as solutions to the planner's problem. Each one of these allocations is Pareto optimal because it solves a planner's problem—again, if it were not Pareto optimal the planner would not choose it. By varying these weights, we can therefore map out many different Pareto optimal allocations or in other words we can trace out the Pareto frontier. Similarly, by varying date-0 wealth in the market economy we can generate different competitive equilibria with different sets of Lagrange multipliers λ_i. By choosing the appropriate distribution of date-0 wealth we can engineer a competitive equilibrium that mimics the solution to the planner's problem with particular Negishi weights and therefore gives rise to a particular Pareto optimal consumption allocation. In this case, the consumers still trade with one another, but we redistribute resources between them to change the distribution of consumption.

This procedure of constructing a competitive equilibrium to deliver a particular Pareto optimal consumption allocation is closely related to the Second Welfare Theorem. The Second Welfare Theorem begins with a Pareto optimal allocation and then gives conditions under which there exists a competitive equilibrium delivering this allocation as an outcome. The conditions involve ensuring that consumers' and firms' maximization problems have well-defined solutions, which in turn in general necessitates assumptions of convexity (e.g., the consumer's utility function is globally concave). These additional assumptions are typically met in our macroeconomic applications so they are not a problem per se, but the statement of a general theorem is more cumbersome so we will not present it here.

Market equilibria, as observed in actual economies, have radically different consumption levels across agents, so viewed from the perspective of frictionless competitive equilibria and

an additive social welfare function of the sort just described, they correspond to points on the Pareto frontier with high Negishi weights on the high-consumption agents. The social welfare function with Negishi weights is just an analytical tool that we can use to construct different Pareto optimal allocations. To be clear, the argument we have made here is not saying that this distribution of consumption is desirable or just but simply that there are some Pareto weights that could make the planner choose that distribution of consumption.

6.3 Inefficient market outcomes

We will now very briefly, by means of simple examples, cover a number of commonly considered departures from the abstract frictionless economy considered above. For each case, we will comment on efficiency properties by describing how the proof of the First Welfare Theorem 6.1 may or may not go through as well as explaining how the more intuitive marginal efficiency conditions may or may not hold. We begin with a case where the culprit is not the market but distortionary taxation and then look at externalities, missing markets, and monopoly power.

6.3.1 Taxes

First, let us consider lump-sum taxes (in a frictionless economy). Lump-sum taxes do not have to be equal across agents; the key is that the amount given to agent i does not depend on the behavior of agent i. From the perspective of the marginal conditions characterizing optima, since the marginal conditions of competitive equilibria do not involve lump-sum taxes or transfers, equilibria remain optimal. From the perspective of our abstract proof, all lump-sum taxes do is redistribute wealth (and thus "utils") across agents, i.e., move us along the Pareto frontier.[3]

Second, let us consider the two most commonly studied taxes in macroeconomic applications: taxes on labor earnings and taxes on capital income. Beginning with taxes on labor earnings, let us consider a tax that is proportional to earnings. Hence, instead of w_t an agent receives $w_t(1-\tau_\ell)$ for each unit of hours worked. In an economy where consumers do not value leisure, this tax does not affect any first-order conditions; hence, it acts as a lump-sum tax and does not disturb the efficiency properties of equilibrium. However, if leisure is valued, the marginal rate of substitution between consumption and leisure will differ from the marginal rate of transformation between consumption and leisure by a factor $1-\tau_\ell$. Similarly, if capital income is taxed, the marginal rate of substitution between goods at t and $t+1$ will differ from the marginal rate of transformation due to the tax.

What goes wrong? First, intuitively the firms and the households face different (after-tax) prices and therefore they may not exploit all of the possible trades that they should. Now looking at our abstract proof of equilibrium efficiency, what goes wrong in trying to use it? The key is that the tax appears in the equations. Suppose that the p in the proof is the

[3]To see this formally, one can consider the government to be one of the consumers in the economy. When we sum across the budget constraints of the consumers in the proof of the First Welfare Theorem, the lump-sum taxes will cancel out as the government's revenue equals the other consumer's tax payments.

pre-tax prices. What would not hold is the point where we say the alternative consumption allocation must cost more: $p\sum_{i\in\mathcal{I}}\tilde{x}_i > p\omega + p\sum_{j\in\mathcal{J}}y_j^\star$. This inequality may not hold because the consumer is maximizing utility with respect to the after-tax prices.[4] Hence the proof does not in general go through. However, the proof is valid when leisure is not valued, because then the leisure chosen, which is an element in x, is zero, so the fact that p is different for this good does not matter. This confirms the intuition that a proportional tax on an inelastic labor supply is non-distortionary.

Similarly, a proportional tax on capital income, e.g., with r_t replaced by $r_t(1-\tau_k)$, will appear in the Euler equation and therefore make the marginal rate of substitution between goods at $t-1$ and t differ from the corresponding marginal rate of transformation. In the proof of the First Welfare Theorem, the relative price of the time t good is higher for consumers than for producers and hence the proof cannot be completed.

6.3.2 Externalities

An externality arises when one agent's activity has payoff relevance to other agents that is not reflected in the market price associated with this activity. Let us use an example with a negative TFP externality: production by one firm damages the production carried out in other firms. Consider a static economy with a representative consumer who values leisure. In particular, the consumer maximizes $u(c,\ell) = \log c - B\ell^{1+1/\theta}/(1+1/\theta)$ subject to

$$c = rk + w\ell$$

by choice of (c,ℓ). Output of a typical firm j equals

$$A(\bar{y})k_j^\alpha \ell_j^{1-\alpha},$$

with $\bar{y} = (\sum_j y_j)/\mu$, where μ is the number of firms. Here, A is a decreasing function, expressing a negative production externality: the higher is total production in the economy, the lower is the productivity of each firm. Let us also assume that μ is large enough that each firm j ignores its own impact on $\bar{y}$; for simplicity, think of the set of firms as a continuum on $[0,1]$, with $\mu = 1$, so that there is a notion of a representative firm. Hence, we can drop the subscript j and the representative firm solves

$$\max_{k,\ell}\ A(\bar{y})k^\alpha\ell^{1-\alpha} - rk - w\ell,$$

thus taking $\bar{y}$ as given. In this static economy, output is determined by equilibrium in the labor market. From the consumer, we obtain

$$\frac{w}{rk+w\ell} = B\ell^{\frac{1}{\theta}}$$

and the firm's first-order conditions imply that

$$\frac{r}{w} = \frac{\alpha}{1-\alpha}\frac{\ell}{k}.$$

[4] We cannot just interpret p as the after-tax price because then when we say firms are profit maximizing, they are not maximizing profits with respect to p.

Combining the two equations, we obtain (with a small amount of algebra) that the equilibrium outcome for hours worked is given by the unique solution to

$$1 - \alpha = B\ell^{1+\frac{1}{\theta}}.$$

A striking feature of this solution is that it does not depend on the strength of the externality as captured by the function A: A does not appear in the equation.[5]

The efficient allocation, on the other hand, is given by the solution to

$$\max_{\ell,y} \ \log y - B\frac{\ell^{1+\frac{1}{\theta}}}{1+\frac{1}{\theta}} \quad \text{subject to} \quad y = A(y)k^{\alpha}\ell^{1-\alpha}.$$

Here, the first-order conditions (with λ denoting the multiplier on the constraint) are

$$B\ell^{\frac{1}{\theta}} = \lambda(1-\alpha)A(y)k^{\alpha}\ell^{-\alpha} \quad \text{and} \quad \frac{1}{y} = \lambda\left(1 - A'(y)k^{\alpha}\ell^{1-\alpha}\right),$$

which delivers

$$\frac{1-\alpha}{1 - A'(y)k^{\alpha}\ell^{1-\alpha}} = B\ell^{1+\frac{1}{\theta}} \quad \text{and} \quad y = A(y)k^{\alpha}\ell^{1-\alpha}$$

as two equations in the two unknowns ℓ and y. The first of these equations can be compared to the equilibrium outcome: the equilibrium is not optimal, unless $A'(y) = 0$, i.e., unless there is no externality. Intuitively, in their input choices, firms do not take into account how their production hurts others, and they should.

What goes wrong? To see how our abstract proof of the First Welfare Theorem would go wrong in this case, first note that an equilibrium with externalities would be defined by letting the Y_js—each firm's production possibility set—be endogenous and interdependent: $Y_j = Y_j((y_{j'})_{j' \neq j})$, where $(y_{j'})_{j' \neq j}$ is the vector of choices of other firms.[6] A key step in our abstract proof was that for each j, $p\tilde{y}_j \leq py_j^{\star}$, from profit maximization. This no longer follows, since the alternative allocation implies a different choice set for firm j, as given by $Y_j = Y_j((y_{j'})_{j' \neq j})$. In concrete terms, if other firms scale down their production relative to that in equilibrium, your choice set improves—your TFP increases—and your original choice was not optimal.[7]

The negative externality considered here is closely related to how the externalities due to climate change are usually modeled. There, the externality is usually assumed to affect TFP, as in our example here. Rather than occurring through overall production, however, the externality occurs through carbon emissions, which result from production of a specific good—energy derived from fossil fuels. Climate change is covered in Chapter 25.

[5] This particular feature, which follows because income and substitution effects cancel in the labor-supply specification, makes the example stark but is not necessary for the arguments here.

[6] For simplicity here we abstract from your own production lowering your own TFP.

[7] In the present example, equilibrium profits are zero, and if all other firms decrease their output below the equilibrium level, your profits can be made positive (and unboundedly large) by simply scaling up your input choices.

6.3.3 Missing markets: an example with constraints on borrowing

A friction that is commonly studied in macroeconomic applications involves "missing markets": restrictions on trade in one way or another. For example, there is no market that allows workers to buy insurance against unfavorable changes in their salaries because of the moral hazard that workers will have incentives to collect the insurance payment rather than work hard. Another example is that households and firms may not be able to borrow much as they would like. Borrowing constraints are thought to be important constraints both for firms (for funding their day-to-day operations as well as long-term investments) and consumers (for purchasing homes and durable goods more generally). These borrowing constraints ultimately stem from features of the economic environment, such as the difficulty of enforcing repayment, but are sometimes modeled as simple constraints. We focus on the case of borrowing constraints here, but the logic applies more generally to settings where certain goods or agreements for borrowing or insurance are not traded.

Let us consider a two-agent dynastic endowment economy: a two-type special case of that presented in Section 5.3.1. The total endowment each period is constant at ω, but let us now assume that agent 1 is endowed with $2\omega/3$ in odd periods and $\omega/3$ in even periods; agent 2 thus has $\omega/3$ and $2\omega/3$ in odd and even periods, respectively (the odd-numbered agent is rich in odd periods). Let us also assume that both agents have logarithmic utility. The economy starts in period 0. In a competitive equilibrium with unrestricted markets we have (i) full consumption smoothing, so that $c_{1,t} = c_1$ for all t and $c_{2,t} = c_2$ for all t; prices for goods at different dates satisfy $p_t = \beta^t$, i.e., the gross real interest rate is $1/\beta$; and (iii)

$$c_1 = \left(\frac{1}{3} + \frac{2\beta}{3}\right)\frac{\omega}{1+\beta}$$

and

$$c_2 = \left(\frac{2}{3} + \frac{\beta}{3}\right)\frac{\omega}{1+\beta}.$$

Consumer 2 is richer, and hence consumes more, due to an endowment stream that is higher in present value because consumer 2 receives the larger endowment one period before consumer 1 does.

Note that the unrestricted competitive equilibrium has active borrowing and lending: agent 1 borrows in even periods and repays in odd periods. Suppose, then, that borrowing is simply not allowed. In terms of a sequence of budget constraints $c_t + q_t a_{t+1} = \omega_t + a_t$, where a is asset holdings and q is price of a one-period real bond delivering 1 unit of consumption next period, no borrowing means that the consumer is facing an additional constraint: $a_{t+1} \geq 0$ for all t. The consumer's maximization problem then reads

$$\max_{\{a_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log\left(\omega_t + a_t - q_t a_{t+1}\right) \quad \text{subject to} \quad a_{t+1} \geq 0 \quad \forall t.$$

Differentiation with respect to a_{t+1} delivers

$$\frac{q_t}{c_t} = \beta \frac{1}{c_{t+1}} + \mu_t,$$

where μ_t is the Kuhn-Tucker multiplier on the borrowing constraint. This multiplier is non-negative: positive when the constraint binds and zero otherwise. Hence, we conclude that under a borrowing constraint, the Euler equation generally is an inequality constraint: $q_t u'(c_t) \geq \beta u'(c_{t+1})$. Intuitively, the marginal value of consumption today can be strictly higher than the consumer could obtain by consumption tomorrow, using market prices, but the consumer cannot increase consumption today further due to the borrowing constraint.

In our simple example, no borrowing means autarky: consumer 1 always consumes $2\omega/3$ in odd periods and $\omega/3$ in even periods, with the remainder of the total endowment consumed by agent 2.[8] The allocation is clearly not optimal, since consumption is not smoothed. What is the prediction for interest rates? In even periods, agent 2 is not constrained so we know that

$$\frac{q_t}{2\omega/3} = \beta \frac{1}{\omega/3}$$

when t is even. Hence, $q_t = 2\beta$: the real interest rate $1/q_t - 1$ is lower than under unrestricted borrowing, since the lender needs to be discouraged from lending.[9] The same logic applies in odd periods: $q_t = 2\beta$. We even see that the real interest rate will be negative provided that $\beta > 0.5$.

What goes wrong? Clearly, marginal rates of substitution are not equalized across agents here. From the perspective of our abstract proof, what goes wrong? As in the case with externalities, the sets from which agents are choosing will change: they will contain more restrictions and they will be endogenous. For example, without borrowing constraints, the main restriction on the consumption possibility sets is non-negativity. With borrowing constraints, it will include an additional constraint for each good: $c_0 \leq a_0 + \omega_0$, $q_0 c_1 \leq \omega_0 + a_0 - c_0 - q_0 \omega_1$, and so on. The equilibrium allocation can be dominated by an alternative allocation—say, one with full smoothing—simply because this alternative allocation is not within the agent's consumption possibility set.

Moreover, as we see here, the consumption possibility sets will now also depend on prices, thus departing conceptually from the abstract setup where the consumption possibility sets are exogenous. As in the externality example, there is an interdependence between the possibility sets the agents face. And just like the externality case, the agents will not take into account how their actions affect the possibility sets that other face. These are *pecuniary externalities*—one agent's choices affect the prices that appear in other agents' constraints. As a result of these pecuniary externalities, if a planner were to select the consumption allocation subject to the budget constraints and borrowing constraints, the planner may choose a different allocation than the competitive equilibrium because the planner would take account of the pecuniary externalities.

[8] As agent 1 is unable to borrow in period 0, agent 2 has nobody to lend to and cannot save. In period 1, the two agents face the same situation as period 0 with their roles reversed.

[9] Values of q_t above 2β will also constitute equilibria, since all agents are then strictly constrained. However, we view the case with an interior solution as the interesting one, since it is the limiting equilibrium for economies with borrowing constraints $a_{t+1} \geq \underline{a}$ where $\underline{a} < 0$, with $\underline{a} \to 0$.

6.3.4 Lack of commitment

In the dynamic equilibrium models we looked at above, we always assumed that all agents engaged in intertemporal trade could commit to delivering on their promises. When such commitment is not available, the allocations are of course affected. One could, for example, imagine that it is possible to default on debt and that such defaults are not punished in any way. This would, given that lenders are rational, lead to de-facto borrowing constraints, as in Section 6.3.3 above. One could also imagine that default can occur but that it is punished, in which case intertemporal trade generally will occur. Note, however, that the punishment mechanism per se needs to be committed to and that it is not clear that punishment will be rational ex post. When punishments are assumed in economic models, there is therefore a presumption that some agent (such as the government) has an ability to commit to it.

Macroeconomic models where consumers cannot fully commit, including where they do default in equilibrium, exist and are interesting cases of "endogenously incomplete markets," but they are not discussed in this text. A more common example involves government policy, such as the case where governments cannot fully commit to future tax policy; this case is discussed in Chapter 15 below. In the context of international economics, another important example is the case where default on sovereign debt (involving lending between countries) can occur; this is a central example in Chapter 24 on emerging markets.

6.3.5 Market power

A well-known case, and also one of great relevance for macroeconomics, is where some agents in the economy have market power. That is, they can, through their own behavior, affect the prices they are transacting at. This may occur on the individual level—you may be able to bargain for a higher salary—or at the firm level—the firm can set the price of its product, or try to bargain with their input suppliers over those prices. We will discuss examples of both of these occurrences in the book. When we study labor-market frictions, it will be natural to discuss bargaining. When we study growth and business cycles, we will study markups, i.e., how firms with market power set prices above marginal cost.

In macroeconomic models with market power, we will almost always assume that this power is quite limited: it is limited to an individual transaction and not affecting aggregates. Thus, for monopoly settings we most often study monopolistic competition, where each firm holds a monopoly over a specific good but faces competition with a continuum of imperfect substitutes, with each one playing a negligible role in the aggregate. Similarly, a worker will be assumed not to have any influence on aggregates in the wage negotiations. Clearly, there are examples where these assumptions are not appropriate (Apple can likely affect the entire market for smartphones, and perhaps even world GDP, in their pricing; particularly gifted vaccine researchers could be seen to have similar powers). But they are likely rare. So let us now briefly and compactly describe the basic, static model of monopolistic competition building on Dixit and Stiglitz (1977).

A model with monopolistic competition

Let us assume a representative consumer with preferences defined over a continuum of imperfectly substitutable goods and labor effort, L:

$$U\left(\left(c(i)\right)_{i=0}^{1}, L\right) = u\left(\left(\int_0^1 c\left(i\right)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}}\right) - v(L) \equiv u(C) - v(L).$$

This function, as we will see, implies a constant elasticity of substitution $\varepsilon \geq 0$ across different consumption goods (a CES function). We will specialize u to be logarithmic below in an example.

The consumer's budget is

$$\int_0^1 p\left(i\right) c\left(i\right) di = y,$$

where y is income. We will now derive a price index by considering how the consumer should allocate goods in the cheapest way in order to reach a given level of C . So consider

$$\min_{\left(c(i)\right)_{i=0}^{1}} \int_0^1 p\left(i\right) c\left(i\right) di \quad \text{subject to} \quad \left(\int_0^1 c\left(i\right)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}} \geq C.$$

We obtain, with λ denoting the multiplier for the constraint,

$$\begin{aligned} p\left(i\right) &= \lambda \frac{\varepsilon}{\varepsilon-1}\left(\int_0^1 c\left(i\right)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}-1}\left(1-\frac{1}{\varepsilon}\right) c\left(i\right)^{-\frac{1}{\varepsilon}} \\ &= \lambda\left(\int_0^1 c\left(i\right)^{1-\frac{1}{\varepsilon}} di\right)^{\frac{\varepsilon}{\varepsilon-1}-1} c\left(i\right)^{-\frac{1}{\varepsilon}} = \lambda\left(\frac{c\left(i\right)}{C}\right)^{-\frac{1}{\varepsilon}}. \end{aligned} \tag{6.4}$$

Multiply by $c(i)$ on both sides, sum across goods, and simplify to obtain

$$\int_0^1 p\left(i\right) c\left(i\right) di = \lambda C.$$

That is, λ can be interpreted as a "unit price of C," the whole basket. What is λ in terms of primitives? Often the multiplier is merely used to set up and solve a maximization problem but sometimes it carries important content, such as here, and then it is relevant to compute its value. To derive a formula for the unit price, use the expression for $p(i)$ again: raise it to $1-\varepsilon$ and sum across i. This delivers

$$P \equiv \lambda = \left(\int_0^1 p\left(i\right)^{1-\varepsilon} di\right)^{\frac{1}{1-\varepsilon}}.$$

So P, the unit price, is itself a CES function that is increasing in all prices and homogeneous of degree one.

A key other implication of equation (6.4) is that it expresses an *inverse demand* function: it expresses a relation between the price of good i and the demand for it, given an overall

level of spending PC. The demand function itself becomes, after solving for $c(i)$ and using $\lambda = P$,

$$c(i) = C\left(\frac{p(i)}{P}\right)^{-\epsilon}.$$

Here we see that the price elasticity of demand is constant and equal to ϵ. (One can replace C by y/P, where y is the consumer's income.)

The demand function is used as a central object in the definition of a monopolistically competitive equilibrium. We will now state it. We assume that one firm produces each kind of consumption good and that all production functions are identical and linear in labor: $c(i) = A\ell(i)$ for all i. We assume that labor supply is exogenous and equal to 1.

A monopolistically competitive equilibrium For the economy described above, a monopolistically competitive equilibrium is a consumption allocation $\{c(i)^\star\}_{\forall i}$, a labor allocation $\{\ell(i)^\star\}_{\forall i}$, a price vector $\{p(i)^\star\}_{\forall i}$, a profit vector $\{\pi(i)^\star\}_{\forall i}$, and a wage $w^\star$ such that

1. $(\{c(i)^\star\}_{\forall i}, L^\star)$, where $L^\star \equiv \int_0^1 \ell^\star(i)di$, solves

$$\max_{(c(i))_{i=0}^1, L} u\left(\left(\int_0^1 c\,(i)^{1-\frac{1}{\varepsilon}}\,di\right)^{\frac{\varepsilon}{\varepsilon-1}}\right) - v(L)$$

   subject to $\int_0^1 p^\star(i)c(i)di = w^\star L + \int_0^1 \pi^\star(i)di$.

2. for each i, $(p^\star(i), c(i)^\star, \ell(i)^\star)$ solves the maximization problem

$$\max_{p,c,\ell}\ pc - w^\star\ell \quad \text{subject to} \quad c = A\ell \quad \text{and} \quad p = P^\star\left(\frac{c}{C^\star}\right)^{-\frac{1}{\epsilon}},$$

   where

$$C^\star = \left(\int_0^1 c^\star(i)^{1-\frac{1}{\varepsilon}}di\right)^{\frac{\varepsilon}{\varepsilon-1}} \text{ and } P^\star = \left(\int_0^1 p^\star\,(i)^{1-\varepsilon}\,di\right)^{\frac{1}{1-\varepsilon}}$$

   and $\pi^\star(i)$ defines the maximum obtained.

Notice (i) that no market-clearing condition is needed as it is satisfied immediately as part of the firms' problems; (ii) that firms make profits, which accrue to the consumer; (iii) that firm i solves a problem that does not depend on i, since all goods are symmetric here (still, we label the solutions with i); and (iv) that a key input into the firm's problem is the inverse demand function, $p^\star(c)$. We stated this last condition with knowledge of form for the inverse demand function for each good; this function, of course, has to be consistent with condition 1 of the definition and this was ensured in our derivations leading up to the definition. A monopolistically competitive equilibrium can thus alternatively be defined to include this demand function explicitly as an equilibrium object, with the added condition that it is consistent with consumer maximization.

To see what implications follow from this setup, let us solve the firm's problem. Substitute the constraints into the objective, so that it reads

$$\max_c\ PC^{\frac{1}{\varepsilon}}c^{1-\frac{1}{\varepsilon}} - \frac{w}{A}c,$$

where we have dropped the i due to symmetry and *s for notational convenience. Clearly, this is a well-defined problem if $\varepsilon > 1$; if $\varepsilon \leq 1$, goods are not sufficiently substitutable, and the monopolist's problem does not have a solution.[10]

$$c = C \cdot \left(\frac{w}{A} \frac{1}{(1-1/\varepsilon)\, P} \right)^{-\varepsilon} \qquad \text{and} \qquad p = \frac{\varepsilon}{\varepsilon - 1} \frac{w}{A};$$

thus, $\mu \equiv \varepsilon/(\varepsilon - 1) > 1$ expresses that the firm charges a *markup* over marginal cost, w/A, that is constant in percent. Profits π are thus given by $(\mu - 1)(w/A)c$.

The equilibrium is symmetric, so $c = C$. We can normalize one price, or a combination of prices, so we set $P = 1$. We then see that the equilibrium wage has to equal $w = A(1 - 1/\varepsilon)$: the wage is below marginal productivity (as $\varepsilon > 1$). Equilibrium work effort and consumption are then solved from $u'(C)w = v'(L)$. With a logarithmic u we obtain

$$A(1 - 1/\varepsilon) = Cv'(C/A),$$

which has a unique solution for C assuming $v(L)$ is convex and, hence, the right-hand side is strictly increasing in C.

It is easy to see that the outcome is inefficient. The planner will choose symmetry across goods and hence simply maximizes $u(C) - v(C/A)$. The outcome is the first-order condition

$$A = Cv'(C/A),$$

whose left-hand side is larger than in the monopolistic case and, hence, output and hours worked are too low in equilibrium.

What goes wrong with market power? Firms choose inputs taking into account how their demand is affected. As a result, firms tend to under-produce relative to the optimum, since that gives them a higher price. Thus, in equilibrium, firms transform labor into consumption goods at a marginal rate that is higher than the rate at which consumers value leisure relative to the consumption good. Our abstract proof of the First Welfare Theorem, moreover, cannot be used directly since an equilibrium with monopoly power is defined quite differently, as we have seen: some agents do not take p as given.

6.3.6 Quantifying welfare losses

In applied macroeconomic analysis, the aim is most often quantitative. The researcher wants to go beyond qualitative statements like "the equilibrium is not optimal" to a quantitative one indicating *how much* worse, or better, one allocation is than another. There are numerous ways to do this and we will focus on the most common one here. We will use distortionary taxes in the context of a representative-agent economy as an example.

So suppose the government has an amount of expenditures—e.g., military purchases—that it needs to finance with taxes. One option, at least in theory, is to use lump-sum taxes.

[10] When $\varepsilon < 1$, infinite profits can be obtained by raising prices toward infinity; the reader is invited to show this by considering the implied maximization problem. The case $\varepsilon = 1$ is special: since revenues are then independent of the price, higher prices, and consequently lower quantities sold and therefore production costs, are always better and no profit maximum exists. Its supremum equals the revenue.

Another one is to tax labor earnings at a proportional rate. How much worse is it to use distortionary taxes? Let us focus on a static model because it is simple; the principles are the same in a dynamic context.

The procedure is simple. First compute an equilibrium for each of the two tax systems. Denote the resulting allocations of consumption and hours worked (c_l, ℓ_l) and (c_d, ℓ_d) (l for lump-sum and d for distortionary). Clearly, $u(c_l, \ell_l) > u(c_d, \ell_d)$, where the utility is defined over consumption and hours (hours affect utility negatively). Now it is always possible, with standard utility functions, to find a $\Delta > 0$ such that $u(c_l(1-\Delta), \ell_l) = u(c_d, \ell_d)$. The Δ expresses how much, in percent, consumption needs to be decreased in the better allocation to generate the same utility as in the worse allocation, while maintaining the same hours choice.

The key in the example is that Δ has a real interpretation; while one could compute the difference in consumer "utils" between two allocations, such a measure would not have an interpretation as "utils" have no meaning per se. There are, of course, alternative ways to define a Δ. One could, for example, imagine both reducing consumption and raising hours worked at the same time. One could also define a Δ as the percentage increase in consumption in the worse allocation that would make it as good as the better allocation. This would deliver a different Δ. One can, finally, define Δ as an amount of wealth that the consumer would need to be given as a lump-sum transfer in order to be indifferent between two allocations.

In dynamic models, the Δ is defined as a percentage change in consumption in all time periods. Consumption need not be constant over time in either of the allocations considered, but a unique Δ can still be defined: it is the percentage decrease in consumption applied in every period that would make the consumer indifferent with the worse allocation.

6.4 Overlapping generations

The efficiency properties in overlapping-generations (OG) models require a separate discussion; they have also been thoroughly studied in the literature. We will emphasize the key results and insights here only. A short summary of the key result is that, first, although there are no frictions—no markets are missing and distortionary taxes, monopoly power, and externalities are absent—equilibria can be, but are not necessarily, Pareto inefficient. Second, there is a simple litmus test that will tell us whether an equilibrium is efficient or not; we will provide it (but not prove it).

Before proceeding, let us note that there are models that share properties with overlapping generations models, such as those where people die randomly and new people appear. Here, equilibria are not necessarily efficient either.

6.4.1 The endowment case

Let us consider a simple example for illustration: there is a representative consumer in each cohort who lives for two periods. Let us also use a concrete, very simple utility function: for every generation $t \geq 0$ preferences are represented by

$$u_t\left(c_y,\, c_o\right) = \log c_y + \log c_o$$

where we evaluate at two arbitrary consumption levels (c_y, c_o). The preferences of generation $t = -1$, who are old as time begins, are similarly represented by $u_{-1}(c) = \log c$.

Let us consider stationary endowment sequences given by

$$\omega_{y,t} = \omega_y$$

and

$$\omega_{o,t} = \omega_o$$

for all t, where t represents the time period; thus, $\omega_{o,t}$ is the endowment of the old (who were born at $t-1$) at time t.

In the competitive equilibrium we consider, trading is sequential and there are no borrowing constraints. With q_t being the price of a bond at t, the agent born at $t \geq 0$ solves

$$\max_{c_y,\, c_o} \ \log c_y + \log c_o$$

subject to

$$c_y + q_t c_o = \omega_y + q_t \omega_o.$$

It is straightforward to solve this maximization problem. It delivers

$$c_{y,t} = \frac{1}{2}\left(\omega_y + q_t \omega_o\right) \tag{6.5}$$

and

$$c_{o,t+1} = \frac{1}{2}\left(\frac{\omega_y}{q_t} + \omega_o\right), \tag{6.6}$$

where we have now given the consumption choices sub-indexes for the time period in which they occur. Note that the consumer's saving when young is $\omega_y - c_{y,t} = \left(\omega_y - q_t \omega_o\right)/2$.

The old agent at time zero maximizes utility subject to the budget $c_{o,0} = \omega_o$ and hence the choice is trivially given by the budget.

Market clearing in this overlapping-generations economy for period $t = 0$ reads

$$c_{o,0} + c_{y,0} = \omega_y + \omega_o.$$

Since the old's choice is given, we conclude that $c_{y,0} = \omega_y$: the young also consumes the endowment. The bond price that clears the market is $q_0 = \omega_y/\omega_o$. It follows that $c_{o,1} = \omega_o$: the old at 1 will consume the endowment in the second period of life as well.

The argument is then repeated and we obtain, period by period, that

$$c_{y,t} = \omega_y,$$

$$c_{o,t} = \omega_o,$$

and

$$q_t = \frac{\omega_y}{\omega_o}.$$

This constant sequence supports the equilibrium where agents do not trade: the prices induce people to consume their initial endowments.

Let us now plug in specific numbers. Let $\omega_y = 3$ and $\omega_o = 1$. It follows that $q_t = 3$. Thus, the gross real interest rate is 1/3; the net interest rate is thus - 67%.

Is this allocation Pareto efficient? Consider the following alternative feasible allocation: for all t,

$$\tilde{c}_{y,t} = 2$$

and

$$\tilde{c}_{o,t} = 2.$$

That is, the alternative allocation $\tilde{c}$ is obtained from a chain of intergenerational goods transfers that consists of the young in every period giving a unit of their endowment to the old in that period. Notice that for all generations $t \geq 0$, this is just a modification of the timing in their consumption, since total goods consumed throughout their lifetime remain at 4. For the initial old, this is an increase from 1 to 2 units of consumption when old. It is clear, then, that the initial old strictly prefer the new allocation. We need to check what the remaining generations think about the change. It is clear that since utility is concave (the log function is concave), this even split of the same total amount will yield a higher utility value: $\log 2 + \log 2 = 2 \cdot \log 2 = \log 4 > \log 3 + \log 1 = \log 3$.

We conclude that the competitive equilibrium, which we solved for uniquely, is not Pareto optimal: it is dominated by an allocation where each cohort gives a transfer when young and receives one when old. But why is the equilibrium not Pareto optimal? There is no friction: no agent is prevented from trading, there are no externalities or elements of market power. We will return to this question shortly, but first let us consider the reverse case: $\omega_y = 1$ and $\omega_o = 3$. Now, q_t becomes 1/3 each period; the net real interest rate is +67%. Again, let us consider the alternative $(2, 2)$ allocation: is it a Pareto improvement? For all generations born at $t \geq 0$, the answer is yes, as in the previous example. However, the old at 0 will be made worse off. Hence, the proposed alternative is not a Pareto improvement. Is there some other, smarter alternative allocation that does the job? The answer is, in fact, no.

In the overlapping generations model, equilibria are sometimes Pareto optimal and sometimes not. We will elaborate on the intuition later, but let us instead go back and revisit our abstract proof of the First Welfare Theorem in Section 6.1. The notation there can, as in our application, include infinite sequences, and the proof goes through in all parts, except possibly in one: when summing the budget over all agents, $\mathcal{I}$ is now infinite. Prices should now be viewed as given by the sequence $p = \{p_0, p_1, p_2, \dots\}$ where $p_t = \prod_{\tau=0}^{t-1} q_\tau$. In our given equilibrium where $q_t = 3$ for all t, then, the present value of the equilibrium allocation is $\omega_o + (\omega_y + \omega_o) + 3(\omega_y + \omega_o) + 3^2(\omega_y + \omega_o) + \dots$ and this sum is infinite. Hence, this proof strategy cannot be used. However, in the equilibrium where $q_t = 1/3$, the present value is finite, and the proof does go through. Thus, we in fact have a proof that there is no allocation that can Pareto improve on the autarkic allocation $(c_{y,t}, c_{o,t}) = (1, 3)$ for all t!

When can equilibria where the present-value budget sum across all agents is infinite be Pareto improved upon? We cannot rely on the standard proof of the First Welfare Theorem, but it turns out that there is a general theorem—one provided in Balasko and Shell (1980)—that gives us the answer. The theorem relies on some assumptions, but these assumptions are rather weak: aside from regularity conditions and a bounded sequence for total endowments, they mainly restrict the curvature of consumers' indifference curves away from the two extreme cases (linear and kinked). The Balasko-Shell result is: A competitive

equilibrium in an endowment economy populated by overlapping generations of agents is Pareto optimal if and only if

$$\sum_{t=0}^{\infty} \frac{1}{p_t} = \infty.$$

The proof is quite involved and we refer the reader to the source. The two special cases we have looked at are consistent with the theorem: $q_t = 3$ for all t implies that $\sum_{t=0}^{\infty}(1/p_t) = \sum_{t=0}^{\infty} 3^{-t} = 3/2$ is finite, and hence the equilibrium is not optimal; $q_t = 1/3$ for all t implies that $\sum_{t=0}^{\infty}(1/p_t) = \sum_{t=0}^{\infty} 3^t$ is infinite, and hence the equilibrium is optimal. Now, however, we can also evaluate the middle case where $(\omega_y, \omega_o) = (2, 2)$. Here, since $q_t = 1$ implies an infinite sum, our abstract proof cannot be used—as in the $(\omega_y, \omega_o) = (3, 1)$ case—but now the theorem tells us that equilibrium is optimal (a infinite sum of 1s is infinite).

The Balasko-Shell theorem can be applied also to non-constant sequences; indeed, it applies also for non-constant endowment sequences. An important observation, however, is that whether $\sum_{t=0}^{\infty}(1/p_t)$ is finite or not does not depend on anything but how p_t behaves as t literally goes to infinity. Thus, whenever the gross real interest rate p_t/p_{t+1} converges, we know that the equilibrium is optimal if and only if the limit net real interest rate is equal to or above 0.[11] Relatedly, if the economy has a finite time horizon, no inefficiency can occur, no matter how the real interest rate evolves: the present value of the sum of all budgets is finite in this case, and the standard proof can be used. It is thus the combination of (i) infinite time and (ii) a corresponding infinite set of cohorts of consumers—each of which has a finite budget—that can make markets fail.

Third and finally, whenever the equilibrium is sub-optimal, a straightforward government policy of transferring resources from the young to the old will allow all generations to be made better off. This is an argument for the introduction of social security as a pure government-mediated transfer scheme: a "pay-as-you-go" system. The government works like a bank here: you give it money when young and get it back when old. What allows the government to achieve a better allocation than markets can deliver? On the one hand, the inefficiency of the overlapping generations model is a pure market failure, and in that sense a government can improve on it. But it does require an ability of the government to implement a sequence of transfers that stretches into eternity. If, in our simple $(\omega_y, \omega_o) = (3, 1)$ case transfers stopped at some point in time, the young in the very last period of transfers will be worse off by having given something away, with nothing in return.

Let us now provide some intuition for the Balasko-Shell result, and let us continue with our example and focus on the "toughest" case: $\omega_y = \omega_o = 2$, where a First Welfare Theorem cannot be proved the standard way and yet applies. First, we restrict attention to *stationary* allocations, i.e., allocations such that $c_{y,t} = c_y$ for all t and $c_{o,t} = c_o$ for all t. So is there a stationary allocation that Pareto dominates (2, 2)? Figure 6.2 shows the resource constraint of the economy, plotted together with the utility level curve corresponding to the allocation (2, 2).

The shaded area is the feasible set; its frontier given by the line $c_y + c_o = 4$. It is clear from the picture with a tangency at (2, 2) (recall that the utility function is $\log c_y + \log c_o$) that it is not possible to find an alternative allocation that Pareto dominates this one. Now,

[11] For an economy where endowments grow at some net rate g in the limit, a similar theorem applies: the equilibrium is optimal if and only if the limit interest rate is equal to or above g.

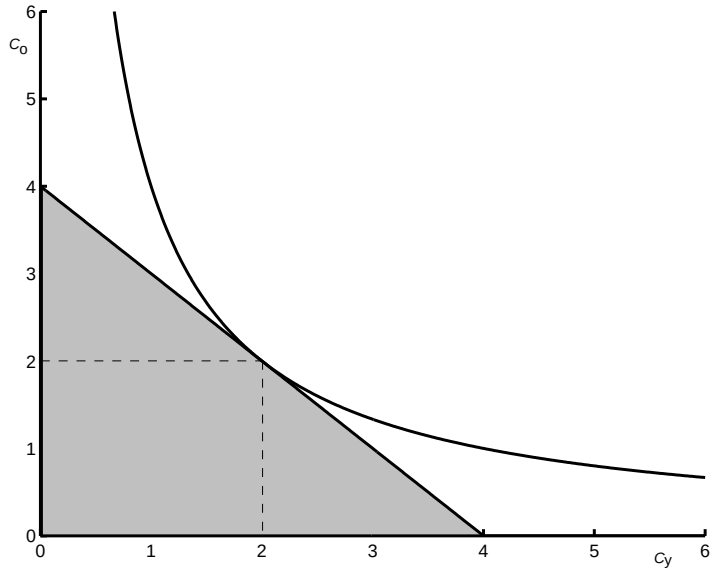


Figure 6.2: Pareto optimality of (2, 2) allocation

let us however admit non-stationary allocations: could there be a non-stationary allocation that dominates (2, 2)? In order to implement such a non-stationary allocation, a chain of inter-generational transfers would require a transfer from young to old at some arbitrary point in time t. The agents giving away endowment units in their youth would have to be compensated when old. The question is how many units of goods would be required for this compensation.

Figure 6.3 illustrates that, given an initial transfer ε_1 from young to old at t, the transfer ε_2 required to compensate generation t must be larger than ε_1, given the convexity of the indifference curves. This in turn will command a still larger ε_3, and so on. Is the sequence $\{\varepsilon_t\}_{t=0}^{\infty}$ thus formed feasible? No: eventually the transfer will exceed the young agent's endowment.

In the "simpler" case, where the equilibrium involves a gross real interest rate less than one, a constant transfer sequence is always possible: one can select another stationary allocation and it is better for everybody. In the case where the gross real interest rate is above one, the argument is as illustrated in Figure 6.3, except even harder, because now the young needs to be compensated even more when old, given that their indifference curves have a higher slope (in absolute value). So no feasible better path exists here either.

6.4.2 Intertemporal production

Intertemporal production in an overlapping generations economy raises further issues. In some cases, the introduction of the possibility to save—in the form of "capital"—can help overcome an inefficiency; in others, it can lead to new market failures. We will keep the discussion brief and merely provide some illustrations.

Let us start with the endowment economy $(\omega_y, \omega_o) = (3, 1)$, where we know that the equilibrium is not optimal. Suppose that there is a simple storage technology allowing the consumers to save $k \geq 0$ units today and receive χk units in return tomorrow. Clearly,

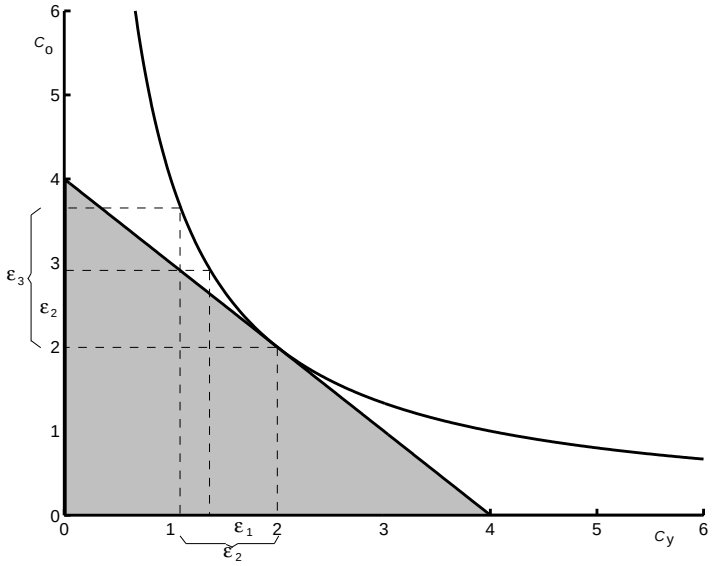


Figure 6.3: (2, 2) cannot be improved upon

if $\chi > 1/3$ they will: it gives them a higher return than in the original equilibrium. Is the resulting equilibrium allocation optimal? We can solve for individuals' optimal saving choices by simply setting $q_t = \chi > 1/3$ in equations (6.5)–(6.6). This is our equilibrium, which again is autarky, but with active individual storage.

If we regard the equilibrium allocation as a new endowment point and ask if Pareto improvements are possible, the answer is given by the Balasko-Shell theorem: improvements are possible if and only if $\chi < 1$. However, this is not the final answer: here, the answer is that the equilibrium is actually optimal when $\chi > 1$ but not when $\chi \leq 1$. In particular, $\chi = 1$ is not optimal. The reason is shockingly simple. In the equilibrium, all young agents store 1 unit and, hence, achieve the consumption allocation $(2, 2)$. So far so good; this is in fact the same allocation for people born at $t = 0$ and later as that with our social-security scheme. However, the old at 0 are still consuming only 1 unit, so a Pareto improvement can be obtained by instead carrying out the social-security scheme: the old at zero are strictly better off and no-one else worse off (everybody else is indifferent).

The market failure with storage here, and $\chi = 1$, is an example of *dynamic inefficiency*: it is possible, by means of different saving choices, to create more resources at at least one point in time without forsaking resources at any other point in time. Thus, there is even a "free lunch" in equilibrium! The key insight here is that *oversaving* can occur in equilibrium. This is different than the market failure with fixed resources that we looked at before in our overlapping generations model: there, it was a matter of redistribution, and here, it is intertemporal production that is inefficient, again despite the absence of frictions.

As we shall see soon, the inefficiency in the special case with a one-for-one storage technology relies on the linearity of the production technology. If the intertemporal production technology is instead neoclassical, i.e., if it has decreasing returns to scale, the Balasko-Shell condition that the equilibrium is efficient if and only if, in the limit, the (gross) real interest rate is greater than or equal to 1, will be recovered. Let us therefore briefly revisit the neoclassical model here. Now let us interpret (ω_y, ω_o) as the endowments of a cohort in labor

efficiency units. An agent's budget sets in period t and $t+1$ are then

$$c_y + s = \omega_y w_t \quad \text{and} \quad c_o = s(1 - \delta + r_{t+1}) + \omega_o w_{t+1},$$

with s denoting saving. If the utility function is strictly quasiconcave, the saving choice at t is given uniquely by some function h:

$$s_t = h\left(w_t,\, r_{t+1},\, w_{t+1}\right)$$

Asset markets clear when $s_t = k_{t+1}$, where k denotes the economy's total capital stock: the young buy the entire capital stock for the next period. Competitive pricing of inputs as usual implies $r_t = F_1\left(k_t,\, \omega_y + \omega_o\right)$ and $w_t = F_2\left(k_t,\, \omega_y + \omega_o\right)$. Thus, our equilibrium can be computed as the solution to the non-linear first-order difference equation

$$k_{t+1} = h\left(F_2(k_t, \omega_y + \omega_o), F_1(k_{t+1}, \omega_y + \omega_o), F_2(k_{t+1}, \omega_y + \omega_o)\right).$$

This equation implicitly determines k_{t+1} as a function of k_t. We then have the following result that allows us to check whether the equilibrium savings choices are dynamically efficient.

Theorem 6.2 *Define $R_t = 1 - \delta + F_1(k_t, \omega_t)$, where ω_t is the total labor endowment at t. Then $\{k_t\}_{t=0}^{\infty}$ is dynamically efficient if and only if*

$$\sum_{t=0}^{\infty} \left[\prod_{s=1}^{t} R_s\left(k_s\right) \right] = \infty.$$

The theorem, whose assumptions are suppressed in the statement for brevity, relies in part on a proof of production efficiency that is similar in spirit to the proof of Pareto efficiency of an overlapping generations allocation with fixed resources.[12] The key point, however, is that it is the same condition as under fixed endowments: $\prod_{s=1}^{t} R_s\left(k_s\right) = 1/p_t$, where p_t is our Arrow-Debreu price of consumption good at t in terms of consumption good at 0.

To obtain intuition, it is instructive to restrict attention to steady states, i.e., solutions to

$$k_{ss} = h\left(F_2(k_{ss}, \omega_y + \omega_o), F_1(k_{ss}, \omega_y + \omega_o), F_2(k_{ss}, \omega_y + \omega_o)\right).$$

Clearly, from the theorem, a steady state is efficient if and only if $R = F_1(k_{ss}, \omega_y + \omega_o) + 1 - \delta \geq 1$, that is, if the net interest rate is non-negative.[13] Let us start with the case of inefficiency. When $F_1(k_{ss}, \omega_y + \omega_o) + 1 - \delta < 1$, an alternative saving plan is possible where savings at time t are reduced by a small amount $\epsilon > 0$. This frees up resources for consumption at t. At $t+1$, there are now fewer resources available, but the reduction is less than ϵ, given that the marginal return to saving was strictly less than one. However, suppose that at $t+1$, saving is again decreased by ϵ relative to the given steady state: then resources are freed up this period as well on net (there is less production, by less than ϵ, but also less saving, by

[12] The assumptions underlying it are that the sequence $\{R_t\}_{t=0}^{\infty}$ be uniformly bounded above and below away from zero and on the production-function curvature being bounded as follows: $0 < a \leq -f_t''\left(k_t\right) \leq M < \infty \;\; \forall t, \forall k_t$, where $f_t(k)$ is defined as $F(k/\omega_t, 1)$ and F is assumed to have CRS.

[13] In a balanced-growth version of this economy, the efficiency requires the interest rate not to be below the growth rate.

ϵ). This procedure is repeated and, as a result, there are strictly more resources available to consume at all points in time beginning with period t. Now suppose there is a steady state with a gross return one or greater than one. If one were to try to create more resources at some point t by reducing saving by ϵ, the resources available at $t+1$ would be reduced by more than ϵ.[14] The future reductions in saving required to keep resources at least as high as before will have to grow over time and will finally, become infeasible. The proofs of these statements, which are available Appendix 6.A, are the key behind understanding the logic of the theorem above.

In conclusion: in the overlapping generations model, equilibria—with or without production—can be efficient or inefficient. The key condition for efficiency, which holds rather generally, is that the asymptotic net real interest rate be non-negative or, in the case of growth, no less than the rate of growth.

6.4.3 Dynamic inefficiency in the warm glow model

The warm glow model, discussed briefly in Chapter 5, can also deliver outcomes with dynamic inefficiency properties. Let us illustrate this point with an example. Assume individuals live for one period only but give bequests according to the warm-glow model, with utility function $u(c_t)+v(b_{t+1})$ for agents alive at time t, and that they are in a standard neoclassical environment; capital accumulation comes from bequests. Thus the agent's budget reads $c_t+b_{t+1} = (1+r_t-\delta)b_t+w_t$. Clearly, the offspring (at $t+1$) obtains b_{t+1} plus the net return.

Assume now that v is linear, $v(b) = Ab$ for some $A > 0$, and that u is continuously differentiable and strictly concave. Then the optimization problem gives a solution for c_t from $u'(c_t) = A$; denote this solution $\bar{c} > 0$. Bequests will then satisfy $b_{t+1} = (1+r_t-\delta)b_t+w_t-\bar{c}$. In equilibrium, $b_{t+1} = k_{t+1}$ so capital accumulation will be given by $k_{t+1} = (1+r_t-\delta)k_t+w_t-\bar{c} = (1-\delta)k_t + F(k_t,1)-\bar{c}$. This equation is a first-order difference equation—a convenience compared to the standard dynastic model—and it is easy to see from simple inspection of the function in (k_t, k_{t+1}) space, that there are two positive steady states. The lowest of these is unstable whereas the highest steady state, $\bar{k}$, is stable. Therefore, for a high enough initial capital stock, there is convergence to the higher steady state. At this steady state, moreover, the slope of the function, $1+F_1(\bar{k},1)-\delta$, is less than 1, but this also means that the long-run real interest rate $r-\delta$, is negative. Hence the economy is dynamically inefficient: there is "too much" capital accumulation, so that output could be increased at all times by lowering the capital stock, along the lines of the previous section. Intuitively, in this model people care about capital savings per se, leading to there being a free lunch in equilibrium.[15]

[14] If the net return is initially zero, it will raise above zero when we reduce saving.

[15] This can occur since the model has a non-standard feature. Note that for a statement about Pareto (in-)efficiency, one would need to think about which objective function to use. Lowering capital below the dynamically inefficient steady state could give all consumers more resources to consume, but they might not be happier given the warm glow from capital itself.

6.5 Optimal government policy

We have now seen a number of examples of how markets may deliver inefficient outcomes. A natural suggestion in each of these cases is to propose a government policy to improve on the allocation. In the case of distortionary taxes, the origin of the inefficiency is in government policy itself, but in the other cases, what would we, as macroeconomic analysts, tell the government?

But let us start with the tax case, because it is still interesting. In particular, it is often argued, and for good reasons, that lump-sum taxes are hard to implement in practice. Thus, tax analysis could be carried out by comparing tax policies that are deemed feasible to implement. For example, in a dynamic model one can compare proportional taxes on labor earnings to proportional taxes on capital income: which is the better system from a welfare perspective, and by how much? Such an approach is referred to as Ramsey analysis, after Frank Ramsey's early work (see Ramsey, 1927). The approach is, however, somewhat problematic since it is often not clear which policies are feasible and which are not. Ramsey taxation will be studied in Chapter 15.

The case of externalities is well understood; here, Pigou (1920) suggested a tax, or transfer, that would counteract the distortion and cancel its effects exactly. In the example of pollution, the externality could be corrected by charging a per-unit output tax on firms equal to what the externality would be, evaluated at the optimal allocation, and it would deliver the optimal allocation as an equilibrium outcome. Similarly, positive externalities can be encouraged with per-unit subsidies where they occur. A different path here would be to follow Coase: assign property rights so that the side effects of agents' actions can be incorporated as market transactions. This path can, potentially, be easier to take, since there is no need for the government to compute what an appropriate tax rate would be: once property rights are assigned and enforced, the property value and the price its owner will charge for its use will be determined by markets and, in the absence of further frictions, lead to an optimal allocation. Sometimes, however, as in the case of damages due to climate change, ownership cannot be assigned: the earth's atmosphere is not possible to own.

Monopoly distortions that result in inefficiently low production are easy to handle in principle: one can, for example, subsidize production at a per-unit rate. Again, this requires a calculation of the appropriate tax rate, which requires much detailed knowledge. Therefore, whenever it is possible, anti-trust regulation can be used to minimize the presence of monopoly pricing. Monopoly power, finally, can play another role in the economy: it can provide incentives to invent new products if, namely, there is patent protection (or it is difficult for other reasons to imitate the product). Invention is typically costly so to incentivize inventors one might wish to exclude others from using an invention. Thus, regulating against monopoly has drawbacks too; we will study this issue in Chapter 13.

6.5.1 Missing markets and the "chicken model"

What about the missing markets case? Recall the endowment economy where endowments alternated between agents and no borrowing was allowed. Here, a reasonably simple policy would seem to be available to the government: each period, tax the agent with a high endowment and transfer the proceeds to the other agent, so as to achieve full consumption

smoothing. Both agents would then be better off, at least if the transfer is small enough. Similarly, in the context of missing insurance markets, the government could simply compensate people who received bad shocks and tax the luckier ones to finance the transfers. Is it reasonable to propose such a policy? This is not so clear. In reality, when a market is missing, it is usually missing for a reason. That reason could, for example, be problems of private information or moral hazard. If borrowing/lending and insurance are so beneficial, why do they not materialize, in the cases where they appear not to be present?[16] A consequence of this point is that it is entirely conceivable that there would be negative side effects of the governments intervention: those side effects that made markets missing in the first place.

The above considerations have given rise to the concept of *a "chicken model" of government*. The model here refers to an argument for government intervention in markets and goes as follows. Assume that (i) people like chicken; (ii) the market economy cannot produce chicken; and (iii) the government can produce chicken. The result then follows: the government should produce chicken. From our perspective, the lesson should be: in cases where government intervention is proposed, think about which friction is at work, and whether it is one that the government is likely to be able to deal with well, or better than the market. Sometimes the answer is likely yes, and other times no.

Clearly one kind of approach available here would be to try to model the causes of frictions, such as private information, explicitly. Thus, analysis following the work of Mirrlees (1971) has been used to study optimal taxes and transfers when markets appear to be functioning imperfectly. Another reason why some markets may not exist is a lack of commitment, as discussed in Section 6.3.4 above.

6.5.2 Redistribution policy

The discussion so far centered around minimizing frictions. In practice, a separate aim is often redistribution, i.e., the idea is not to Pareto improve on the given allocation but rather to achieve a more equitable distribution of consumption even if some agents are made worse off. In macroeconomic models with heterogeneity, some of which will be studied later in this book, researchers often adopt a social welfare function to guide policy choices. In such cases, the welfare weights on different agents would represent the policymaker's preferences but, of course, not necessarily those of the researcher.

An often used social welfare function is an additive ("utilitarian") formulation. The most common assumption then is that the utilities of the agents are weighted equally. Clearly, equal weights would amount to equal consumption in an economy where direct, non-distortionary redistribution is available. Hence, equal weights embody a strong desire for equality. If taxes are distortionary, then equal weights still express the same desire but the optimal level of redistribution will typically not fully eliminate consumption inequality.

An argument for equally-weighted utilitarian social welfare functions that have been used is the "behind-the-veil-of-ignorance" notion. So imagine that a person, before they are born into a household somewhere in the world, possibly also without knowing what

[16] Think about whether you should write an insurance contract with your fellow graduate students, making sure that your post-graduation salaries are all the same, after taking transfers between you into account.

genetic skills they will have once born, is asked to consider potential distributional policies. Then redistribution could potentially be viewed as an optimal insurance scheme; in concrete terms, maximize $\pi u(c_A) + (1-\pi)u(c_B)$ subject to a resource constraint $\pi c_A + (1-\pi)c_B = \pi\omega_A + (1-\pi)\omega_B$, where $\omega_A > \omega_B$ would be the market incomes of the two types of agents; π is the fraction of people who will be born as A types. Clearly, this optimization problem embodies equal weights and delivers $c_A = c_B$. The insurance solution cannot be offered by markets, since the agents are not around to sign the contract before they are born, but governments can nevertheless carry out a policy which achieves the insurance outcome. Is this therefore a chicken model of government? Not so much; it is more seen a potential guiding philosophical principle with which you may agree or disagree.

Chapter 7

Uncertainty

Many aspects of economic life are not fully predictable. For example, at the aggregate level, technologies and policies can change in unpredictable ways. Life is even more uncertain at the individual level due to the unpredictability of income, health, and other events. We now introduce the main techniques that macroeconomists use to incorporate uncertainty into our analysis of the economy.

This chapter covers several issues. First, we introduce analytical tools that are helpful in analyzing stochastic economies. These tools include mathematical concepts related to stochastic processes as well as economic concepts related to decision making under uncertainty. We then use these tools to analyze the social planner's problem for a version of the neoclassical growth model with stochastic productivity. This important example is introduced in Section 7.3. We then discuss how agents trade with each other in an uncertain economic environment and present the competitive equilibrium of the stochastic neoclassical growth model. Finally, we briefly present an environment with incomplete insurance markets.

7.1 Stochastic processes

A **stochastic process** is a collection of random variables indexed by time. Suppose at each date t there is a random outcome X_t. The collection of these random variables $\{X_t : t \in \mathcal{T}\}$ is a stochastic process, where the set $\mathcal{T}$ is the span of time we are interested in. When modeling an economy with uncertainty, we typically assume that some fundamental features of the economy follow an exogenous stochastic process. For example, we often assume that the level of productivity fluctuates over time and is modeled as an exogenous stochastic process. This randomness in fundamentals generates randomness in endogenous variables. Our economic model determines the stochastic process these endogenous variables follow. We will now introduce some general properties of stochastic processes before turning to a few of the main types of processes used by macroeconomists.

7.1.1 Properties of stochastic processes

When working with stochastic processes we often need to take expectations of them. One way to think about the expectation of a random variable X_t that is part of a stochastic process is to imagine multiple realizations of the entire stochastic process $\{X_t^{(i)} : t \in \mathcal{T}, i \in 1, \cdots, I\}$, where i indexes the different realizations. Taking the expectation across i (i.e., add using probability weights) then gives the unconditional expectation. A related concept is a conditional expectation given information up to some point in time. For example we could suppose we have observed the realization of the stochastic process up to date t and then imagine different possible realizations for $t+1$ and later dates. Taking the expectation of X_{t+1} across these different realizations is then an expectation conditional on information through date t. We often use $\mathbb{E}_t[X_{t+1}]$ to denote such a conditional expectation. To be clear, $\mathbb{E}_t[X_{t+1}]$ means the expectation of X_{t+1} conditional on all information available at date t, not just the realization of X_t observed at date t. A very useful property of expectations is the **law of iterated expectations**, which for any dates $t < s < \tau$ says

$$\mathbb{E}_t\left[\mathbb{E}_s\left[X_\tau\right]\right] = \mathbb{E}_t\left[X_\tau\right].$$

A proof of the law of iterated expectations appears in the appendix.

In addition to expectations, we are often interested in the second moments of a stochastic process. These second moments are captured by the autocovariances. The j-th autocovariance is given by $\mathbb{E}\left[(X_t - \mu_t)(X_{t-j} - \mu_{t-j})\right]$ where μ_t is the unconditional expectation of X_t.

Many economic theories assume or imply stochastic processes that are stationary. A process is **covariance stationary** if neither the unconditional expectations nor the autocovariances depend on time t. In practice, stationarity means that the consequences of a shock to the process eventually fade. If this were not the case, as time goes by, the effects of all the shocks that have occurred would accumulate and the distribution of X_t would change, e.g., become more and more dispersed, as t increases.

In most cases we look at, a stationary process will be **ergodic**. For an ergodic process, observing a long time series allows us to understand the distribution of the stochastic process. For example, the unconditional expectation of X_t is the expectation across different possible realizations of X_t. For an ergodic process, an average of a long time series $(1/T)\sum_{t=1}^{T} X_t$ will converge to $\mathbb{E}[X_t]$ as $T \to \infty$.[1] Ergodicity is useful because in practice we often have data on a single long time series.

A stochastic process is (first-order) **Markov** if its current value summarizes all the available information that is useful for predicting its future realizations. For example, the conditional distribution of X_{t+1} conditional X_t is the same as the conditional distribution of X_{t+1} conditional on $\{X_\tau\}_{\tau \leq t}$, in which case knowledge of X_τ for $\tau < t$ does not add any relevant information beyond that contained in X_t. More formally, a Markov process satisfies $\Pr[X_{t+k} = x | X_\tau \forall \tau \leq t] = \Pr[X_{t+k} = x | X_t]$ $\forall t, k \geq 0$. In dynamic programming, it is convenient if a single state variable summarizes the process. For this reason, macroeconomists often work with Markov processes.

Over time, a deterministic sequence might converge to a particular value. A stochastic process, on the other hand, might continually be subject to random shocks and therefore not

[1] See Hamilton (1994) for the conditions under which a stationary process is ergodic.

settle down to a particular value but nevertheless there is a sense in which it can converge. A stationary distribution (or invariant distribution) $\bar{\pi}(X)$ of a Markov process X has the property that if X_t has distribution $\bar{\pi}(X)$ then X_{t+1} also has distribution $\bar{\pi}(X)$.

7.1.2 Markov chains

Let x_t be a random variable that takes on values in the discrete set $\mathcal{X} = \{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_N\}$. If x_t follows a Markov process, the probability that a particular x_{t+1} occurs depends only on x_t and not on earlier values. A Markov process with a discrete state space is called a **Markov chain**. As x_t takes on discrete values, we can summarize the process in terms of a **transition matrix**. Suppose the probability of moving from state i to j is given by P_{ij}. We can then collect these transition probabilities in a matrix P. Row i represents the probabilities of states 1 through N occurring next period conditional on the current state being i. As exactly one of these states will occur, these probabilities must sum to one. By this logic, each row of P must sum to one.

In addition to the transition matrix, a full description of the stochastic process also requires knowing the initial distribution over the states. We will represent this distribution as a $1 \times N$ vector of probabilities π_0 such that the ith element of π_0 gives $\Pr[x_0 = \bar{x}_i]$.

Given the initial distribution π_0, the transition matrix determines the probability distributions for all x_t for $t \geq 1$. We have

$$\Pr[x_1 = \bar{x}_j] = \sum_{i=1}^{N} \Pr[x_1 = \bar{x}_j | x_0 = \bar{x}_i] \times \Pr[x_0 = \bar{x}_i] = \sum_{i=1}^{N} P_{ij} \times [\pi_0]_i$$

where $[\pi_0]_i$ is the ith element of π_0. Notice that this sum is the product of π_0 against the ith column of P. Repeating this logic for each $j = 1, \cdots, N$ we have $\pi_1 = \pi_0 \times P$. This relationship generalizes to $\pi_{t+1} = \pi_t \times P$ and by repeated substitution we have

$$\pi_{t+k} = \pi_t \times P^k.$$

This is a very useful property of Markov chains: the conditional distributions k periods ahead can be found by raising the transition matrix to the power k.

For a Markov chain, a stationary distribution $\bar{\pi}$ is one for which $\bar{\pi} = \bar{\pi} \times P$. If we transpose this definition we have $\bar{\pi}' = P' \times \bar{\pi}'$ and we can see that the stationary distribution is the eigenvector of P' associated with a unit eigenvalue.[2]

To give an example with an economic interpretation, suppose workers can be employed or unemployed. Unemployed workers find jobs with probability $f \in (0, 1)$ and lose (or separate from) jobs with probability $s \in (0, 1)$. Then we can represent the transitions across the employed/unemployed states by the transition matrix

$$P = \begin{pmatrix} 1-s & s \\ f & 1-f \end{pmatrix},$$

[2] A transition matrix will always have a unit eigenvalue. As the rows of P sum to one, we know $\mathbf{1} = P\mathbf{1}$, where $\mathbf{1}$ is a column vector of ones. This equation says P has a unit eigenvalue and the eigenvalues of P and P' are the same.

where the first state represents being employed and the second represents being unemployed. As the stationary distribution has two probabilities that sum to one, there is only one unknown. Let $\bar{u}$ be the stationary (steady state) unemployment rate so that we have $\bar{\pi} = \begin{bmatrix} 1-\bar{u} & \bar{u} \end{bmatrix}$. The eigenvector of P' associated with a unit eigenvalue solves

$$\begin{pmatrix} -s & f \\ s & -f \end{pmatrix} \begin{pmatrix} 1-\bar{u} \\ \bar{u} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Or, equivalently,

$$\bar{u} = s(1-\bar{u}) + (1-f)\bar{u}.$$

The steady state unemployment rate is equal to the mass of employed who separate plus the mass of unemployed who remain unemployed. Solving this equation yields

$$\bar{u} = \frac{s}{s+f}.$$

Notice that there is a unique stationary distribution of this economy. Suppose we start our economy with an unemployment rate u_0. As time goes by, it is also straightforward to check that the unemployment rate will converge to $\bar{u}$ regardless of what u_0 we start with.[3]

When will a Markov chain more generally converge to a unique stationary distribution? It is not always the case, as the following examples demonstrate. Consider a Markov chain with a transition matrix equal to the identity matrix. No matter what initial distribution we start with, the distribution will forever remain stationary so there is not a unique distribution although it converges immediately. As another example, consider the Markov chain with transition matrix given by

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This Markov chain has a unique stationary distribution of [1/2, 1/2], but unless we start with that distribution, the Markov chain will never converge to it. To see this, suppose we start with probability mass $p \neq 1/2$ on the first state and $1-p$ on the second. The mass on the first state will oscillate between p and $1-p$ forever.

A simple necessary condition for convergence and uniqueness is that from each state there is a positive probability of moving to any other state. This condition is actually substantially stronger than we need. Weaker conditions are as follows. State j of a Markov chain is said to be reachable from state i if there is some n such that $\Pr(x_n = \bar{x}_j | x_0 = \bar{x}_i) > 0$. A Markov chain is said to be irreducible if all states are reachable from all other states. State i of a Markov chain is said to be aperiodic if there is some n such that for all $n' \geq n$ $\Pr(x_{n'} = \bar{x}_i | x_n = \bar{x}_i) > 0$. A Markov chain is aperiodic if all states are aperiodic. An irreducible Markov chain has a unique stationary distribution, $\bar{\pi}$, and if it is aperiodic then $\lim_{t\to\infty} \pi_0 P^t = \bar{\pi}$ for all π_0.

[3] We obtain $u_{t+1} = s(1-u_t) + (1-f)u_t = s + (1-f-s)u_t$. By repeated substitution we see that $u_t = s(1+\lambda+\lambda^2+\cdots+\lambda^t u_0)$, with $\lambda \equiv 1-f-s$, which will converge to $s/(s+f)$ since both s and f are strictly between 0 and 1.

7.1.3 Autoregressive processes

We now turn our attention to stochastic processes with continuous distributions. A very common formulation is the autoregressive process of order one or **AR(1) process** for short. The stochastic process x_t follows an AR(1) if it satisfies

$$x_t = \rho x_{t-1} + b\varepsilon_t + (1-\rho)\mu \tag{7.1}$$

where $x_t \in \mathbb{R}$, ρ, b, and μ are scalar coefficients, and ε_t is a stochastic process that satisfies $\mathbb{E}_{t-1}[\varepsilon_t] = 0$, $\mathbb{E}_{t-1}[\varepsilon_t^2] = 1$, and $\mathbb{E}_{t-1}[\varepsilon_t \varepsilon_{t+s}] = 0$ for all $s > 0$.[4] We call the process (7.1) an AR(1) because x_t depends on just one lagged value x_{t-1}.

We can calculate moments of this process by expressing x_t as a moving average of past ε's. Through repeated substitution we arrive at

$$\begin{aligned} x_t &= \mu + b\varepsilon_t + \rho b\varepsilon_{t-1} + \rho^2 b\varepsilon_{t-2} + \cdots \\ &= \mu + b\sum_{s=0}^{\infty} \rho^s \varepsilon_{t-s}. \end{aligned}$$

If $|\rho| < 1$, the effect on x_t of shocks, ε, in the distant past vanishes and the process is stationary. Taking an unconditional expectation we see $\mathbb{E}[x_t] = \mu$ since $\mathbb{E}[\varepsilon_{t-s}] = \mathbb{E}[\mathbb{E}_{t-s-1}[\varepsilon_{t-s}]] = 0$ for all s. The unconditional variance of x_t is given by

$$\operatorname{Var}[x_t] = \sum_{s=0}^{\infty} (b\rho^s)^2 \operatorname{Var}[\varepsilon_{t-s}] = \frac{b^2}{1-\rho^2},$$

where we have used the fact that the ε's have unit standard deviation and are uncorrelated across time. Similarly, the covariance of x_t and x_{t+j} is

$$\begin{aligned} \operatorname{Cov}(x_t, x_{t+j}) &= \mathbb{E}[(x_t - \mu)(x_{t+j} - \mu)] \\ &= \mathbb{E}\left[\left(b\sum_{s=j}^{\infty} \rho^{s-j}\varepsilon_{t-s+j}\right)\left(b\sum_{s=0}^{\infty} \rho^s \varepsilon_{t+j-s}\right)\right] \\ &= b^2\left(\rho^j + \rho^{j+2} + \rho^{j+4} + \cdots\right) \\ &= \frac{b^2\rho^j}{1-\rho^2}. \end{aligned}$$

The correlation between x_t and x_{t+j} is therefore ρ^j. In summary, the parameter μ determines the level of the process, the parameter b determines the volatility of the process, and the parameter ρ determines the persistence of the process.

7.1.4 Linear stochastic difference equations

We now will generalize our autoregressive specification to allow for vector-valued random variables. Let x_t be a column vector in $\mathbb{R}^n$. Let ε_t be a random variable in $\mathbb{R}^m$. The stochastic process ε_t is assumed to satisfy

$$\mathbb{E}_t[\varepsilon_{t+1}] = 0, \tag{7.2}$$

[4] The assumption that ε_t has unit variance is a normalization as the parameter b scales the effects of ε_t on x_t.

$$\mathbb{E}_t\left[\varepsilon_{t+1}\varepsilon_{t+1}'\right] = I, \tag{7.3}$$

and

$$\mathbb{E}_t\left[\varepsilon_{t+1}\varepsilon_{t+s}'\right] = 0 \quad \forall s > 1. \tag{7.4}$$

The second condition says the elements of ε_t are uncorrelated with each other and have unit standard deviation. The third condition states that ε_t is uncorrelated across time. We assume that x_t follows a linear stochastic difference equation:

$$x_t = Ax_{t-1} + B\varepsilon_t + C. \tag{7.5}$$

The $n \times n$ matrix A controls how x_{t-1} affects x_t. If all the eigenvalues of A are smaller than 1 in absolute value, then the effects of past shocks will eventually fade and x_t will be a stationary process. The $n \times m$ matrix B captures the effects of ε_t on x_t.[5] Lastly, the $n \times 1$ vector C affects the mean of x_t as we describe next.

The unconditional expectation of x_t is

$$\mu \equiv \mathbb{E}\left[x_t\right] = A\mathbb{E}\left[x_{t-1}\right] + C = A\mu + C.$$

Solving this equation yields $\mu = (I - A)^{-1}C$. Similarly, let $\Gamma(0)$ be the unconditional covariance matrix of x_t. The definition of a covariance matrix gives us

$$\begin{aligned}\Gamma(0) &= \mathbb{E}\left[(x_t - \mu)(x_t - \mu)'\right] \\ &= \mathbb{E}\left[A(x_{t-1} - \mu)(x_{t-1} - \mu)'A' + B\epsilon_t\epsilon_t'B'\right] \\ &= A\Gamma(0)A' + BB',\end{aligned} \tag{7.6}$$

where we have used the fact that ε_t is independent of x_{t-1}.[6]

We are often interested in the behavior of a stochastic process following a particular event. For example, we might be interested in how the economy would behave following a TFP shock. In this thought experiment, we suppose no further shocks occur. Using $C = \mu - A\mu$, rewrite (7.5) as

$$x_t - \mu = A(x_{t-1} - \mu) + B\varepsilon_t.$$

Suppose there is a particular shock ε_t at t and then no future shocks. Repeated substitution yields

$$x_{t+h} - \mu = A^{h+1}(x_{t-1} - \mu) + A^h B\varepsilon_t.$$

The effect of ε_t on x_{t+h} is given by $A^h B\varepsilon_t$. This change in the future evolution of the process is called the impulse response of x_t to the shock ε_t. The function $\mathcal{F}(h) = A^h B\varepsilon$ is the **impulse response function** of x to the particular shock ε. The impulse response function tells us how x responds to the shock as a function of the time since the shock has occurred. We saw some examples of impulse responses in the deterministic, non-linear Solow model in Section 3.5.2. If the Solow model is extended to include stochastic shocks, the results in this section would apply only to a linear approximation to the Solow model.

[5] The assumption reflected in (7.3) that the ε_t are uncorrelated with each other and have unit standard deviations is a normalization since the matrix B can rescale their standard deviations and impart correlations across their effects.

[6] Equation (7.6) is a Lyapunov equation and can be used to solve for $\Gamma(0)$. Many software packages are available to solve such equations.

7.2 Choice under uncertainty

We will now begin our discussion of how agents make choices under uncertainty. In this section we start by introducing a framework for modeling uncertainty in a way that can introduce uncertainty without restricting ourselves to a specific stochastic process. We then discuss preferences over risky consumption outcomes. Risk aversion is an important aspect of preferences in an uncertain environment and we will demonstrate the implications of risk aversion through a portfolio choice problem.

7.2.1 Stochastic events

To incorporate uncertainty into economic theory, it is often convenient to define a stochastic event that determines all the risky outcomes. The idea here is that there are many different ways the world may take shape in the future and our uncertainty is that we do not know which of these "worlds" we live in. As our theories are typically dynamic, we need to allow for our uncertainty to resolve over time. Let $\omega_t \in \Omega_t$ be the stochastic event realized at date t and let $\omega^t = \{\omega_0, \omega_1, \cdots, \omega_t\} \in \Omega^t$ be the history of events up to date t. To give an example, suppose your income each month can either be high or low. The event ω_t determines whether your income is high or low in month t and ω^t gives a list of all the past events from which we can infer your past incomes. Figure 7.1 shows an example of how these stochastic events could unfold for $t = 0, 1, 2$ when there are two possible realizations of ω_t at each date: $\omega_t \in \{0, 1\}$.

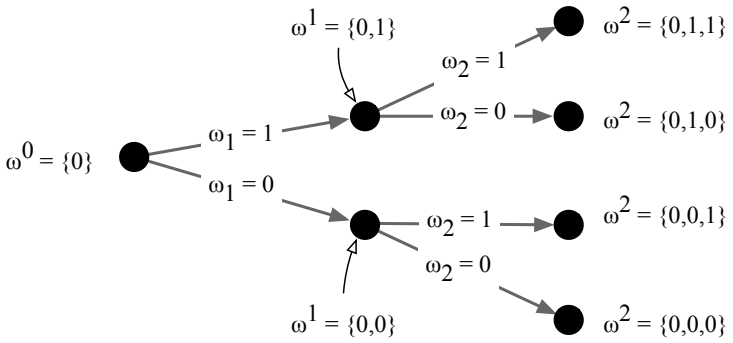


Figure 7.1: Example event tree.

The probability that history ω^t will be realized at date t from the perspective of date 0 is given by $\pi_t(\omega^t)$. The conditional probability of ω^t given ω^τ for $t > \tau$ is $\pi_t(\omega^t|\omega^\tau)$. An outcome at date t is a function of the history up to date t. For example, the balance in your bank account reflects not just the randomness in your current income, but also the fluctuations in your previous incomes (as well as the changes in spending they induce). We could write your assets as $a_t(\omega^t)$ to indicate that it depends on the whole history of events leading up to date t.

7.2.2 Expected utility and risk aversion

We now extend the preferences we introduced in Chapter 4 to incorporate uncertainty according to *expected utility theory*. Thus, utility over stochastic events is then a convex linear combination of a function $u(c)$, where c is random, with the linear coefficients being the probabilities with which the different outcomes for c are realized. Applied to the context of preferences over time, suppose $\{c_t(\omega^t) : \forall t, \omega^t\}$ and $\{\tilde{c}_t(\omega^t) : \forall t, \omega^t\}$ are two consumption processes. We will say the c process is preferred to the $\tilde{c}$ process if and only if

$$\sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u(c_t(\omega^t)) > \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u(\tilde{c}_t(\omega^t)).$$

In many situations, we will not write the stochastic events explicitly because the time subscript on the variables is sufficient to keep track of which histories they depend on. Using that notational convention, the above statement would be

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t) > \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(\tilde{c}_t).$$

As we get started, however, it is useful to be explicit about the histories.

Risk aversion is a fundamental concept of choice under uncertainty that will allow us to explain, among other things, why consumers buy insurance and why savers hold diversified portfolios. Risk aversion is the idea that a less risky consumption stream is preferred to a more risky stream with the same expected consumption. Mathematically, risk aversion is implied by Jensen's inequality when $u(\cdot)$ is concave. If $u(\cdot)$ is linear, then the consumer ranks consumption streams by the expected level of consumption and is said to be risk neutral.

The curvature of $u(\cdot)$ determines the level of risk aversion. The coefficient of absolute risk aversion is defined as $-u''(c)/u'(c)$. A more concave utility function leads to higher risk aversion. The second derivative is normalized by the first derivative to capture the change in curvature as utility changes.[7] The coefficient of absolute risk aversion refers to the attitude towards changes in consumption of a given (absolute) size. The constant absolute risk aversion (CARA) utility function is given by $u(c) = -\exp(-\alpha c)$. For this utility function, the coefficient of absolute risk aversion is α at all levels of consumption.

Alternatively, the coefficient of relative risk aversion measures a consumer's attitude to proportional changes in consumption. It is defined as $-cu''(c)/u'(c)$. Our power utility function

$$u(c) = \frac{c^{1-\sigma} - 1}{1 - \sigma},$$

motivated by its consistency with exact balanced growth, has a constant coefficient of relative risk aversion. Taking derivatives, we see the coefficient of relative risk aversion is simply σ. As the coefficient of relative risk aversion is the same at all levels of consumption, these preferences are also known as the constant relative risk aversion (CRRA) utility function.

[7] In expected utility theory, a positive affine transformation of a utility function represents the same preferences. That is, the utility function $v(c)$ defined by $au(c) + b$ for constants $a > 0$ and b is equivalent to the utility function u. Normalizing the coefficient of risk aversion by $u'(c)$ makes the coefficient of absolute risk aversion invariant to such an affine transformation of u.

In most macroeconomic applications, the CRRA function is used because, as we have discussed, it is the only one consistent with balanced, long-run growth. From the perspective here, using a CRRA function means that decisions about risks today and one hundred years ago—when the level of consumption was much lower—would have been made the same way if the risks were the same in percentage terms.

Recall from Section 4.2.4 that the elasticity of intertemporal substitution is the elasticity of consumption growth with respect to the real interest rate. More precisely, the consumption Euler equation for power utility implies $d\log[c_{t+1}/c_t]/dR_{t+1} = 1/\sigma$, thus implying that the elasticity of intertemporal substitution actually equals the inverse of the coefficient of relative risk aversion. The intuition here is that both of these features of the utility function are determined by the curvature of the utility function. If the utility function displays strong diminishing marginal utility, then a consumer will be unwilling to accept higher consumption in one state of the world in exchange for low consumption in another state of the world (i.e., risk aversion is high). Similarly, the consumer will be unwilling to accept low consumption in period t in exchange for high consumption in period $t+1$ (i.e., is unwilling to substitute intertemporally). Intuitively put, one single parameter guides the desire for smoothing consumption across time and states of nature.[8]

7.2.3 Portfolio choice

A simple portfolio choice problem can help illustrate the differences between the two concepts of risk aversion introduced above. Suppose an individual has $W > 0$ units of wealth at date 0 to allocate between a risk-free asset and a risky asset. The assets will pay off at date 1 and the individual will consume the proceeds. The (gross) return on the risk-free asset is known to be R^f while the return on the risky asset is unknown and denoted Z. Let A be the assets invested in the risky asset while $W - A$ are invested in the risk-free asset. The investor's decision problem is

$$\max_A \mathbb{E}_0 \left[u\left(R^f(W-A) + ZA \right) \right].$$

The first-order condition of this problem is

$$\mathbb{E}_0 \left[u'\left(R^f(W-A) + ZA \right) \left(Z - R^f \right) \right] = 0.$$

Suppose the investor's utility function is the CRRA utility function $u(c) = c^{1-\sigma}/(1-\sigma)$. The first-order condition of the portfolio choice problem becomes

$$\mathbb{E}_0 \left[\left(R^f(W-A) + ZA \right)^{-\sigma} \left(Z - R^f \right) \right] = 0.$$

Rearranging we arrive at

$$\mathbb{E}_0 \left[\left(R^f \left(1 - \frac{A}{W} \right) + Z\frac{A}{W} \right)^{-\sigma} \left(Z - R^f \right) \right] = 0,$$

[8] There is a generalization of the power-function preferences that allows one to separate risk aversion from intertemporal substitution using two separate parameters. This case will be discussed in Chapter 16 below.

where we have brought $W^{-\sigma}$ outside the expectation because it is known at date 0. This equation determines A/W as a function of σ, R^f, and the distribution of the risky asset return. Notice that the solution for A/W does not depend on W, which means that at any level of wealth, the investor will allocate the same fraction of savings to risky assets: the rich and the poor choose the same risk exposure.

Now suppose the investor has the CARA utility function $u(c) = -\exp(-\alpha c)$. The first-order condition then becomes

$$\mathbb{E}_0\left[\alpha \exp\left\{-\alpha R^f W\right\} \exp\left\{-\alpha\left(Z - R^f\right) A\right\}\left(Z - R^f\right)\right] = 0.$$

We can bring $\exp\left\{-\alpha R^f W\right\}$ outside the expectation because it is known at date 0 to arrive at

$$\mathbb{E}_0\left[\exp\left\{-\alpha\left(Z - R^f\right) A\right\}\left(Z - R^f\right)\right] = 0.$$

This equation gives a solution for A that does not depend on W. In this case, the investor allocates a particular level of savings (an absolute number of goods or dollars) to risky assets regardless of their wealth. This means that the rich have a lower risky share than do the poor.

In the data, as we shall see in Chapter 21, the rich on average choose a higher risky share. Neither of these simple models can match this fact but the CRRA case is more in line with the data.

7.3 The stochastic growth model

We can now use the tools of choice under uncertainty and stochastic processes to analyze a stochastic version of the neoclassical growth model. As briefly discussed in Chapter 3 above, in this model, total factor productivity (TFP) is assumed to follow an exogenous stochastic process. The fluctuations in TFP then give rise to endogenous fluctuations in output, consumption, and investment; the model was first studied, as a planning problem, in Brock and Mirman (1972), and then became a workhorse framework macroeconomics, and we devote Chapter 14 below to it. Through this important example, we will discuss how our methods for dynamic optimization can be extended to allow for uncertainty.

7.3.1 A two-period economy

To begin, suppose the economy exists for two periods. In period 0, the level of TFP in period 1 is not known. Let ω_1 be the stochastic event in period 1. Let $\pi_1(\omega_1)$ be the probability of ω_1. TFP at date 1 is given by $A_1(\omega_1)$.

The economy is inhabited by a representative household with expected utility preferences that ranks consumption streams according to

$$U = u(C_0) + \beta \sum_{\omega_1 \in \Omega_1} \pi_1(\omega_1) u\left(C_1(\omega_1)\right), \tag{7.7}$$

where $C_1(\omega_1)$ is the level of consumption if ω_1 occurs. In period 0, the economy is endowed with K_0 units of capital, which are used to produce $Y_0 = K_0^\alpha$ units of output. This output is then used for consumption and investment subject to the date-0 resource constraint

$$K_1 + C_0 = K_0^\alpha + (1-\delta)K_0. \tag{7.8}$$

In period 1, the value of ω_1 becomes known and the economy produces $Y_1(\omega_1) = A_1(\omega_1)K_1^\alpha$. As there are no further periods, there is no reason to invest in capital so the resource constraint in period 1 is

$$C_1(\omega_1) = A_1(\omega_1)K_1^\alpha + (1-\delta)K_1. \tag{7.9}$$

The first important thing to note is that A_1, Y_1, and C_1 are all functions of the event ω_1. From the perspective of date 0, agents do not know A_1 and therefore they cannot know how much will be produced or consumed. When we formulate a decision problem in date 0, the agents will not choose specific values for Y_1 and C_1, but rather they will choose a plan for how they will respond to each realization of ω_1. This is an important feature of optimization under uncertainty: the choice variable is a contingent plan for actions following each history of stochastic events.

The planner's problem for this economy is to choose C_0, K_1, and $\{C_1(\omega_1) : \forall\omega_1\}$ to maximize (7.7) subject to (7.8) and (7.9). We can form the Lagrangian as

$$\begin{aligned}\mathcal{L} = u(C_0) + \beta \sum_{\omega_1\in\Omega_1} \pi_1(\omega_1) u\left(C_1(\omega_1)\right) - \lambda_0\left[K_1 + C_0 - K_0^\alpha - (1-\delta)K_0\right] \\ - \sum_{\omega_1\in\Omega_1} \lambda_1(\omega_1)\left[C_1(\omega_1) - A_1(\omega_1)K_1^\alpha - (1-\delta)K_1\right],\end{aligned}$$

where λ_0 and the λ_1's are Lagrange multipliers on (7.8) and (7.9), respectively. As (7.9) must hold for each realization of ω_1 we treat that as a separate constraint for each ω_1. We therefore have separate Lagrange multipliers for each ω_1 and we use the sum to include all of them in the Lagrangian.

Taking the first-order conditions for this problem, we have

$$u'(C_0) = \lambda_0,$$

$$\lambda_0 = \sum_{\omega_1\in\Omega_1} \lambda_1(\omega_1)\left(\alpha A_1(\omega_1)K_1^{\alpha-1} + 1 - \delta\right),$$

and

$$\beta\pi(\omega_1)u'(C_1(\omega_1)) = \lambda_1(\omega_1). \qquad \forall\omega_1.$$

The second line is the first-order condition for K_1. On the right-hand side we have a sum over all possible realizations of ω_1 because when the planner chooses K_1 they do not know which ω_1 will occur so they need to take into account how K_1 affects output and consumption after each one. In contrast, the third line is the first-order condition with respect to $C_1(\omega_1)$ for a specific ω_1 and there is one such equation for each ω_1.

Combining the first-order conditions to eliminate the Lagrange multipliers we have

$$u'(C_0) = \beta \sum_{\omega_1 \in \Omega_1} \pi(\omega_1) u'\left(C_1(\omega_1)\right) \left(\alpha A_1(\omega_1) K_1^{\alpha-1} + 1 - \delta\right).$$

This is the stochastic consumption Euler equation for the planner. The left-hand side is the marginal utility loss in date 0 from saving one more unit. The right-hand side is the expected marginal utility gain in period 1 from saving one more unit. Notice that we sum over ω_1 and weight the outcomes by $\pi_1(\omega_1)$ so we are taking an expectation. The term $\alpha A_1(\omega_1) K_1^{\alpha-1} + 1 - \delta$ is the marginal increase in resources from increasing K_1 and $u'\left(C_1(\omega_1)\right)$ is the marginal utility of consuming more. The former is the return on saving, which is stochastic as it depends on TFP. The latter reflects the fact that the return on capital is valued differently after different realizations of ω_1. This is due to diminishing marginal utility—when TFP is high, consumption will be high and the marginal value of consuming more is low. The uncertainty in TFP generates uncertainty in $C_1(\omega_1)$, which in turn generates uncertainty in marginal utility.

The Euler equation brings us to another important point about dynamic optimization under uncertainty that turns out to be general: the return to saving is evaluated differently in different states of the world. We do not just focus on the expected return, but instead a weighted average that accounts for the different value of resources in different situations.

7.3.2 An infinite-horizon economy

We now consider an infinite-horizon version of the model. At each date t there is a realization ω_t and the date-0 probability of a history ω^t is given by $\pi_t(\omega^t)$. The representative household has preferences given by

$$U = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u\left(C_t(\omega^t)\right). \tag{7.10}$$

Unlike the two-period model, consumption at date t now depends on the whole history of stochastic events. At each date, a stochastic TFP is realized and used to produce output

$$Y_t(\omega^t) = A_t(\omega^t) F(K_t(\omega^{t-1}), L_t(\omega^t)),$$

where $L_t(\omega^t)$ is the labor input and $A_t(\omega^t)$ is total factor productivity. The production function is twice continuously differentiable in K and L, is strictly increasing and strictly concave in both arguments, and is constant returns to scale. Note that the capital that is used in production at date t, denoted K_t, is selected at date $t-1$ and therefore can only depend on the information that is available at the time it is selected. Therefore K_t is a function of ω^{t-1}, not ω^t.

The economy is endowed with one unit of labor each period and we assume, for simplicity, that there is no preference for leisure so that labor supply is inelastic and equal to one. The aggregate resource constraint at date t is

$$K_{t+1}(\omega^t) + C_t(\omega^t) = f\left(A_t(\omega^t), K_t(\omega^{t-1})\right), \tag{7.11}$$

where we have defined $f(A,K) \equiv AF(K,1)+(1-\delta)K$. The economy begins with an initial endowment of capital, K_0. Negative capital holdings are not possible.

The planner's problem for this economy is to maximize (7.10) subject to (7.11) where the constraint applies to each t and each ω^t. The Lagrangian of this problem is

$$\mathcal{L} = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \left\{ \pi_t(\omega^t)\beta^t u\left(C_t(\omega^t)\right) - \lambda_t(\omega^t)\left[K_{t+1}(\omega^t) + C_t(\omega^t) - f\left(A_t(\omega^t), K_t(\omega^{t-1})\right)\right]\right\}.$$

When we take the first-order conditions of this problem a key point is that our choice of $K_{t+1}(\omega^t)$ will affect production in $t+1$ for all histories ω^{t+1} that are possible given that we have already reached ω^t. For example, refer back to Figure 7.1 and suppose we have reached $\omega^1 = \{0,0\}$ at date 1 and we are choosing $K_2(\omega^1)$. This choice of capital will affect production at date 2 for histories $\{0,0,0\}$ and $\{0,0,1\}$ because these are possible following ω^1. This choice of capital will not affect production for histories $\{0,1,0\}$ or $\{0,1,1\}$ because these are not possible given ω^1. The first-order condition for $K_{t+1}(\omega^t)$ is therefore

$$\lambda_t(\omega^t) = \sum_{\{\omega^{t+1}|\omega^t\}} \lambda_{t+1}(\omega^{t+1}) f_2\left(A_{t+1}(\omega^{t+1}), K_{t+1}(\omega^t)\right), \tag{7.12}$$

where the notation $\{\omega^{t+1}|\omega^t\}$ indicates that we sum over the histories ω^{t+1} that are possible given ω^t and $f_i(\cdot,\cdot)$ represents the partial derivative with respect to ith argument. The first-order condition with respect to $C_t(\omega^t)$ is

$$\pi_t(\omega^t)\beta^t u'\left(C_t(\omega^t)\right) = \lambda_t(\omega^t).$$

Using this to eliminate the Lagrange multipliers in (7.12) we arrive at the consumption Euler equation for this problem

$$u'\left(C_t(\omega^t)\right) = \beta \sum_{\{\omega^{t+1}|\omega^t\}} \pi_{t+1}(\omega^{t+1}|\omega^t) u'\left(C_{t+1}(\omega^{t+1})\right) f_2\left(A_{t+1}(\omega^{t+1}), K_{t+1}(\omega^t)\right), \tag{7.13}$$

where we have defined the conditional probability $\pi_{t+1}(\omega^{t+1}|\omega^t) = \pi_{t+1}(\omega^{t+1})/\pi_t(\omega^t)$. The consumption Euler equation has the same interpretation as in the two-period economy. The right-hand side has a weighted average of the returns on savings with weights corresponding to the marginal utility of consumption in different states of the world.

In many applications, the time subscripts on variables are sufficient to indicate the history of events they depend on in which case we can rewrite (7.13) as

$$u'(C_t) = \beta \mathbb{E}_t\left[u'(C_{t+1}) f_2(A_{t+1}, K_{t+1})\right]. \tag{7.14}$$

The $\mathbb{E}_t$ indicates we are taking conditional expectation just as our sum over $\{\omega^{t+1}|\omega^t\}$ does.

We can now use equation (7.11), at (t,ω^t) and at all the nodes in the following period, to eliminate consumption from the Euler equation (7.13). In the deterministic model, we followed the corresponding method and arrived at a second-order difference equation in capital. Here, we obtain a second-order *stochastic* difference equation in capital, which holds at all nodes in the event tree. In the deterministic model, we also had a transversality condition—as a second-order difference equation and an initial value for capital leave one

degree of freedom to choose capital—and we could then pin down a solution to the difference equation. With uncertainty, the situation has much of the same structure: a transversality condition must be added to determine a solution, and it has the same interpretation as before: it is a self-imposed constraint not to over-accumulate at infinity, expressed as an *expected* present value.[9]

Clearly, solving for a stochastic sequence of capital levels appears daunting. Closed-form solutions exist in special cases, but only under very special assumptions; one is when the utility function is logarithmic, the production function is Cobb-Douglas, and $\delta = 1$. For that case it is possible to verify that a constant saving rate is optimal. In other cases, one must apply numerical methods to solve the model. One such approach is to linearize the model around the steady state of the deterministic model; we show how to do this in Section 7.3.4. If non-linearities are believed to be important, a way forward is to solve the model numerically in a dynamic-programming version of the model. We now look at how it is formulated.

7.3.3 A recursive formulation

We will now analyze a recursive version of the same economy. To do so, we will assume that TFP follows a first-order Markov process so we only need to keep track of the most recent realization in order to know the distribution of its future realizations. As recursive modeling keeps track of the history of the economy explicitly through well chosen state variables, it is customary not to use histories of stochastic events (ω^t) in this context. Following this convention, we let $\pi(A'|A)$ be the probability of A' occurring next period given the current TFP A. The planner's problem can now be expressed as the following Bellman equation

$$V(A,K) = \max_{C,K' \geq 0} \left\{ u(C) + \beta \sum_{A'} \left[\pi(A'|A) V(A',K') \right] \right\}$$

subject to

$$K' + C = f(A,K).$$

The difference compared to dynamic programming under certainty is that now the Bellman equation has an expectation over continuation values. In the next period, there will be a value of having states (A',K') but A' is not known yet. As in the case with certainty, the recursive formulation delivers the same solution as the sequential formulation of the problem. It may seem surprising that this works out since here we are only taking expectations one period into the future, while in the sequential formulation we take expectations of outcomes far in the future. However, the two sets of expectations are actually the same. In the recursive formulation $V(A',K')$ is a random variable that includes a term $\mathbb{E}[V(A'',K'')|A']$. By the law of iterated expectations, the expectation of this term conditional on A becomes $\mathbb{E}[V(A'',K'')|A]$. The same logic applies for the value function at all future dates.

[9] In the case with stochastic shocks, there is still just one degree of freedom given the first-order conditions. To see this intuitively, the Euler equation at any node (t,ω^t) can be used to solve for $k_{t+1}(\omega^t)$ as a function of $k_t(\omega^{t-1})$ and an expression involving future values $k_{t+2}(\omega^t,\omega_{t+1})$ and substituted into the previous Euler equation. When repeated infinitely many times, we have a first equation involving k_0 and $k_1(\omega_0)$ only—the latter is the only remaining degree of freedom.

To analyze the recursive economy, we can proceed with the same steps as we would for the recursive economy without uncertainty. We can substitute the constraint in to the Bellman equation

$$V(A,K) = \max_{K'} \left\{ u\left(f(A,K) - K'\right) + \beta \sum_{A'} \left[\pi(A'|A) V(A',K')\right] \right\}$$

and take the first-order condition with respect to K' to obtain

$$u'(C) = \beta \sum_{A'} \pi(A'|A) V_2(A',K'). \tag{7.15}$$

The envelope condition gives us

$$V_2(A,K) = u'(C) f_2(A,K).$$

Using this to eliminate the derivative of the value function in (7.15) we obtain a similar consumption Euler equation as we had before

$$u'(C) = \beta \sum_{A'} \pi(A'|A) \left[u'(C') f_2(A,K)\right].$$

We can then rewrite this Euler equation as a functional equation that determines the savings policy function. To do so, let $g(A,K)$ denote the choice of K' as a function of states (A,K). Then write the resource constraint as $C = f(A,K) - g(A,K)$. Substituting these definitions into the Euler equation we have

$$u'\left(f\left(A,K\right) - g\left(A,K\right)\right) = \beta \sum_{A'} \pi(A'|A) \left[u'\left(\underbrace{f(A', g(A,K)) - g(A', g(A,K))}_{=C'} \right) f_2(A', g(A,K)) \right]. \tag{7.16}$$

This equation must hold for all (A,K) and it implicitly defines the function $g(A,K)$ that is the solution to the planner's problem. In the next section, we will use this functional Euler equation to derive a complete solution to the planner's problem.

7.3.4 Solving the model via linearization

Section 7.3.3 derived a functional Euler equation that the solution to the planner's problem must satisfy. In this model, productivity follows an exogenous stochastic process. We will now show how we can use a linear approximation to the functional Euler equation to derive a linear stochastic difference equation that the endogenous variables in the model must follow. We will then use the properties of linear stochastic difference equations from Section 7.1.4 to derive properties of the planner's solution.

We will now assume that TFP follows an AR(1) process given by $A' = \rho A + (1-\rho)\bar{A} + \varepsilon'$ with $\mathbb{E}_t[\varepsilon'] = 0$. We will approximate the behavior of the economy around the deterministic steady state, which is the same notion of a steady state we have studied before. Now that we have shocks in the model, the interpretation of the steady state changes. The economy

will never converge to the steady state if it is constantly hit by shocks.[10] The deterministic steady state is the point the economy would converge to if all shocks take their unconditional expectation forever and the agents in the model expect this.[11]

We thus take a linear approximation of (7.16) around the steady state. For a variable X we will use the notation $\hat{X}_t \equiv X_t - \bar{X}$, where $\bar{X}$ is the steady state value.[12] Our linear approximation (which is tedious but straightforward to derive) is

$$(f_K - g_K)\hat{K} + (f_A - g_A)\hat{A} = \mathbb{E}\left[\begin{array}{c} f_K\left((f_K - g_K)\left(g_K\hat{K} + g_A\hat{A}\right) + (f_A - g_A)\hat{A}'\right) \\ +\frac{u'}{u''}\left(f_{KK}g_K\hat{K} + f_{KK}g_A\hat{A} + f_{KA}\hat{A}'\right) \end{array} \middle| A\right],$$

where f_K is the derivative of $f(A,K)$ evaluated at the steady state. Other abbreviations f_i and g_i $(i = A, K)$ represent corresponding derivatives. f_{ij} $(i,j = A,K)$ are second derivatives, also evaluated at the steady state. Notice that $\hat{A}' = \rho\hat{A} + \varepsilon'$ so the distribution of A' given A is determined by the distribution of ε. We will write the expectation as summing over ε' and substitute in for A'. Because this is a linear equation, we can pass the expectation operator inside the right-hand side to obtain

$$(f_K - g_K)\hat{K} + (f_A - g_A)\hat{A} = \beta\left[\begin{array}{c} f_K\left((f_K - g_K)\left(g_K\hat{K} + g_A\hat{A}\right) + (f_A - g_A)(\rho\hat{A} + \mathbb{E}[\varepsilon'])\right) \\ +\frac{u'}{u''}\left(f_{KK}g_K\hat{K} + f_{KK}g_A\hat{A} + f_{KA}(\rho\hat{A} + \mathbb{E}[\varepsilon'])\right) \end{array}\right].$$

As $\mathbb{E}[\varepsilon'] = 0$, it is as if there is no uncertainty and $\hat{A}'$ is treated as if it is known to be at its expected value $\rho\hat{A}$. This is a general feature of analyzing a stochastic economy through linearization known as **certainty equivalence**—once one linearizes the economy, only expected values matter. In particular, the variance of the exogenous shock does not influence outcomes.

Recall that equation (7.16) must hold for all values of A and K. The equation above is a linear approximation to (7.16) and must hold for each $\hat{K}$ and each $\hat{A}$. The only way this can be true is if the coefficients on $\hat{K}$ on the left-hand side equal those on $\hat{K}$ on the right-hand side and, similarly, the coefficients on $\hat{A}$ match. Imposing that these coefficients match gives us two equations that allow us to solve for g_K and g_A. Conveniently, one equation contains g_K only. Therefore, starting with the coefficients on $\hat{K}$ we have

$$(f_K - g_K) = \beta\left[f_K(f_K - g_K)g_K + \frac{u'}{u''}f_{KK}g_K\right].$$

Rearrange to obtain

$$g_K^2 - \left[1 + \beta^{-1} + \frac{u'}{u''}\frac{f_{KK}}{f_K}\right]g_K + \beta^{-1} = 0. \tag{7.17}$$

[10] In addition, the average value of variables will not coincide with those at the deterministic steady state if the model is non-linear. As the shock variance becomes smaller and smaller, however, they will become increasingly similar and, in the limit where the shock variance is zero, coincide.

[11] That the agents perceive the environment to be deterministic is a subtle but important point. There is an alternative notion of a steady state in which the agents in the model perceive there to be risk but ex post all the shocks are realized at their unconditional means. Such a steady state is sometimes called a "stochastic steady state" or a "risky steady state."

[12] Previously in the text, we used this notation for deviations in logs; both linear and log-linear deviations are used in practice.

where we have used $\beta f_K = 1$, which follows from the steady state Euler equation. Equation (7.17) is a quadratic equation in g_K. The equation will have one root less than one and one root greater than one. To verify this, note that the coefficient on g_K^2 is positive so (7.17) is an upward facing parabola as shown in shown in Figure 7.2; the quadratic intersects the y-axis at β^{-1}, which is positive; and at $g_K = 1$ the quadratic takes the value $-\frac{u'}{u''}\frac{f_{KK}}{f_K}$, which is negative if the utility and production functions are both strictly increasing and strictly concave. The quadratic therefore takes the form shown in Figure 7.2 and has one root between 0 and 1 and one greater than 1. The relevant root is the smaller one, because it is

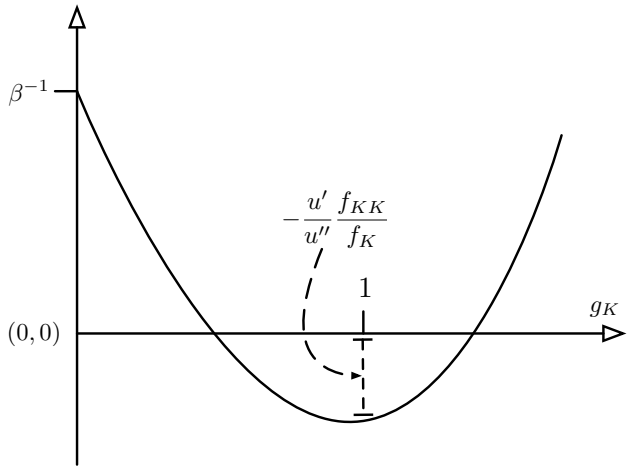


Figure 7.2: Quadratic equation to determine g_K in the linearized stochastic growth model.

the one that is consistent with the transversality condition. The root above one is explosive in that any initial condition or shock has an increasing effect on the economy.

Solving for g_A is more straightforward because matching coefficients on g_A leads to a linear relationship with the solution

$$g_A = \left(f_K - g_K + 1 - \rho + \frac{u'}{u''}\frac{f_{KK}}{f_K} \right)^{-1} \left[(1-\rho) f_A - \rho \frac{u'}{u''}\frac{f_{KA}}{f_K} \right].$$

We have solved for the derivatives of the savings policy rule. The level of the policy rule is determined by the requirement that $(\bar{K}, \bar{A})$ is a steady state so we have

$$K' = \bar{K} + g_K(K - \bar{K}) + g_A(A - \bar{A}). \tag{7.18}$$

If we augment this equation with

$$A' = \bar{A} + \rho(A - \bar{A}) + \varepsilon', \tag{7.19}$$

we have a system of two linear stochastic difference equations and we can apply the techniques described in Section 7.1.4 to analyze the behavior of the economy.

For a numerical illustration of the properties of the stochastic growth model we will make some specific assumptions about the production function and the parameters of the model. We assume $f(A, K) = AK^\alpha + (1-\delta)K$, with $\alpha = 0.3$, $\delta = 0.02$. The persistence of TFP is $\rho = 0.95$, the standard deviation of the innovations is 0.5%, and $\bar{A} = 1$. Finally, we assume $u(c) = \log(c)$ and $\beta = 0.99$.

With our parameterized solution to the model we can generate random draws of $\{\varepsilon_t\}_{t=0}^{T}$ and iterate equations (7.18) and (7.19) forward to simulate the behavior of the economy. Notice here that we are simulating the behavior of the state variables. Given the state variables it is straightforward to calculate the simulated path for output (using the production function) and the simulated path for consumption (using the aggregate resource constraint). Simulated paths for these variables are shown in Figure 7.3(a). Notice how TFP and output exhibit much more high-frequency variation than do capital and consumption. Consumption is smooth because of the diminishing marginal utility of consumption—when output is high, it is preferable to save some of the extra resources to consume them later rather than consume all of them when marginal utility is low. Capital is smooth because it reflects the accumulation of savings over many periods—one period of high investment will not make a big percentage difference to the capital stock. Panel (b) of Figure 7.3 shows the impulse response functions following a one standard deviation shock to TFP. K_t is pre-determined so it does not respond in the period the shock occurs. Therefore, on impact of the shock, output increases by the same percentage amount as TFP. Consumption increases, but not as much as output as some of the increase in output is directed to increased savings. Over time, capital increases and the increase in output exceeds that of productivity.

Figure 7.3(c) shows the results from simulating the economy for a long time and forming a histogram with the simulated data on the capital stock. The solid line in the figure shows the theoretical unconditional distribution of the capital stock as calculated from equation (7.6).[13] As the figure shows, the economy fluctuates in the vicinity of the steady state. Sometimes capital drifts higher, sometimes lower, but it tends to return towards the steady state level. Panel (d) of the figure shows why this is the case. The figure plots $g(A, K)$ as a function of K for two levels of A—one high and one low. The dashed line is the 45-degree line. For low K and high A, the savings policy rule is above the 45-degree line and the capital stock will increase. Similarly, for high K and low A, the savings policy rule is below the 45-degree line and the capital stock will decrease. As A fluctuates, the savings policy will shift up and down leading the capital stock to fluctuate. But note that for high A and high K, or low A and low A, the savings policy intersects the 45-degree line. These intersections imply that capital will not move out of this range (unless A gets even higher or even lower).

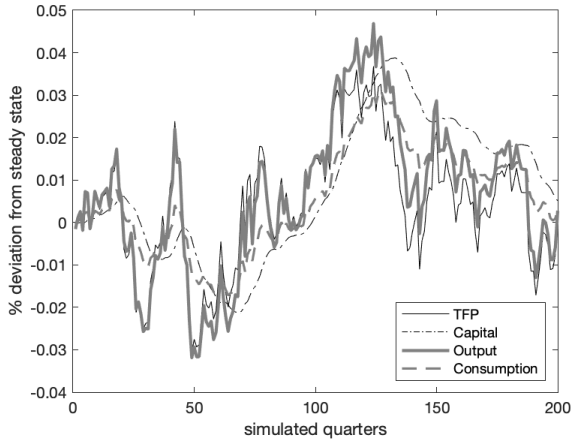
7.4 Competitive market trade under uncertainty

We will now begin to discuss market interactions in an uncertain economy and in the next section we will use these theoretical tools to analyze a decentralized equilibrium of the stochastic growth model. Before we get to that, we will first discuss how models of trade under uncertainty allow us to analyze the way agents insure themselves by sharing risks between them.

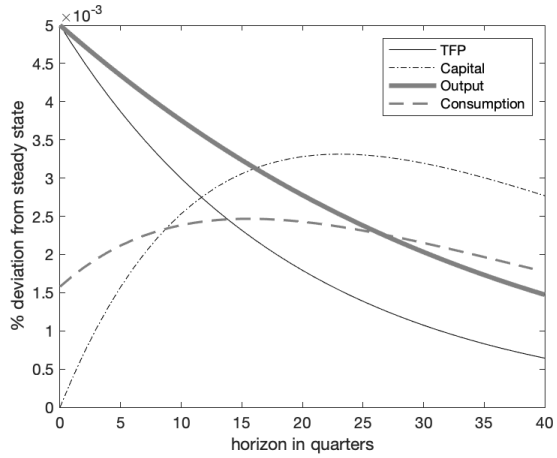
Broadly speaking, we can classify economic models of trade under uncertainty into two groups: complete markets models and incomplete markets models. In models with complete

[13]Specifically, we simulated the economy using Gaussian random variables for ε. As the dynamics of the economy are linear, the distribution of the state variables is also Gaussian. We use (7.6) to solve for the unconditional covariance matrix A and K. We have plotted a Gaussian distribution with the unconditional variance of K and a mean equal to the steady state capital stock.

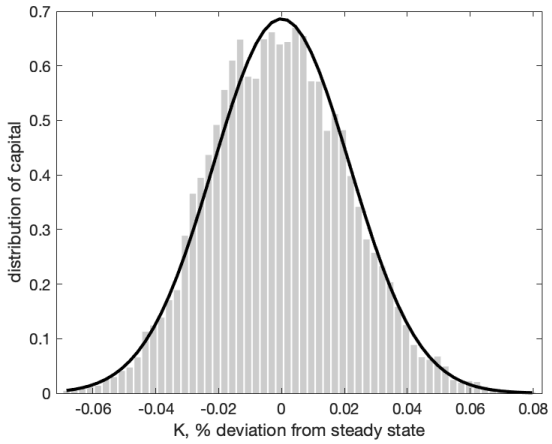
(a) Simulated sample paths



(b) Impulse response functions



(c) Distribution of K



(d) $g(K, A)$ for two levels of A

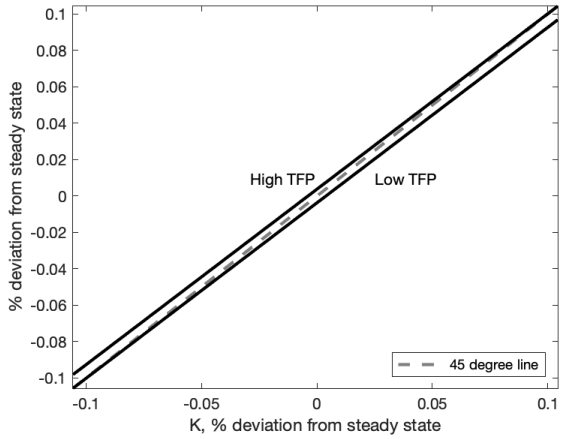


Figure 7.3: Numerical illustration of the stochastic growth model.

markets, agents can buy and sell goods with contracts tailored to every possible state of the world. They can write a contract for how they will behave after every possible history ω^t. This means that any possible risk can be insured at some price. In incomplete markets models, some of these contracts are not available—some risks can not be insured against. In this chapter, we will mostly focus on complete markets models. We start here because it is simpler, not because it is more realistic, but we will introduce an incomplete markets environment in Section 7.6.

Suppose the economy is populated by a set $\mathcal{I}$ of infinitely-lived households; this set could be the continuum [0,1], as in most of Chapter 5, or it could be a different, perhaps smaller set. Each household has expected utility preferences given by

$$\sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \pi_t(\omega^t) \beta^t u\left(c_{i,t}(\omega^t)\right),$$

where $c_{i,t}(\omega^t)$ is the consumption of household $i \in \mathcal{I}$ after history $\omega^t \in \Omega^t$. We do not spell out the nature of ω_t just yet, as it depends on the population details (the nature of the set

$\mathcal{I}$). We assume u is strictly increasing and strictly concave. Each household is endowed with a stochastic income stream that depends on the stochastic event. In particular let the income of household i at date t after history ω^t be $y_{i,t}(\omega^t)$. The good is perishable, meaning that it cannot be stored from one period to the next and must be consumed in the period it arrives in the economy. A consumption allocation is feasible if the total consumption of goods at t and ω^t is equal to the total endowment of goods

$$\sum_{i\in\mathcal{I}} c_{i,t}(\omega^t) \leq \sum_{i\in\mathcal{I}} y_{i,t}(\omega^t).$$

We will begin by considering a market structure in which there is trade only at date 0. For each date t and history ω^t, there is a contract that says the seller will pay the buyer one unit of good at that date if that history has been realized, and nothing otherwise. These are called *Arrow securities*. We will denote the date-0 price of obtaining one unit of goods after history ω^t as $p_t(\omega^t)$. This price is denominated in terms of date-0 consumption.

The decision problem of household i is to choose a contingent plan of $\{c_{i,t}(\omega^t) : \forall t, \omega^t\}$ to maximize the expected utility preferences subject to the budget constraint

$$\sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} p_t(\omega^t)c_{i,t}(\omega^t) = \sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} p_t(\omega^t)y_{i,t}(\omega^t).$$

This budget constraint says the date-0 cost of the consumption plan is less than the date-0 value of the income process. It is as if the agent, at date 0, sells claims to all their future income and then uses those funds to buy consumption goods to be delivered at future dates if particular histories occur.

Formally, we have the following.

Definition 13 *An **Arrow-Debreu competitive equilibrium** is a set of stochastic sequences $\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}$, for each $i \in \mathcal{I}$, and $\{p_t(\omega^t) : \forall t, \omega^t\}$ such that*

1. *for each i, $\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}$ solves*

$$\max_{\{c_t(\omega^t):\forall t,\omega^t\}} \sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} \beta^t\pi(\omega^t)u(c_t(\omega^t)) \quad \text{subject to} \quad \sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} p_t(\omega^t)c_{i,t}(\omega^t) = \sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} p_t(\omega^t)y_{i,t}(\omega^t)$$

2. *$\sum_{i\in\mathcal{I}} c^\star_{i,t}(\omega^t)di = \sum_{i\in\mathcal{I}} y_{i,t}(\omega^t)di$ for all (t, ω^t).*

To characterize the equilibrium, the Lagrangian of household i is

$$\mathcal{L} = \sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} \beta^t\pi_t(\omega^t)u(c_{i,t}(\omega^t)) + \lambda_i\left[\sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t} p_t(\omega^t)\left(y_{i,t}(\omega^t) - c_{i,t}(\omega^t)\right)\right],$$

where λ_i is the Lagrange multiplier on the date-0 budget constraint. The first-order condition of household i with respect to $c_{i,t}(\omega^t)$ is

$$\beta^t\pi_t(\omega^t)u'(c_{i,t}(\omega^t)) = \lambda_i p_t(\omega^t) \tag{7.20}$$

We can understand several properties of the equilibrium consumption allocation from the first-order condition.

Insurance Consider equation (7.20) for two different values of ω^t, call them ω^t and $(\omega^t)'$ and take the ratio of these two equations to arrive at

$$\frac{\beta^t \pi_t(\omega^t) u'(c_{i,t}(\omega^t))}{\beta^t \pi_t((\omega^t)') u'(c_{i,t}((\omega^t)'))} = \frac{\lambda_i p_t(\omega^t)}{\lambda_i p_t((\omega^t)')}.$$

If the prices satisfy $p_t(\omega^t) = \bar{p}_t \times \pi_t(\omega^t)$ with $\bar{p}_t > 0$, which we call **actuarially fair prices**, we have

$$u'(c_{i,t}(\omega^t)) = u'(c_{i,t}((\omega^t)'))$$

and

$$c_{i,t}(\omega^t) = c_{i,t}((\omega^t)'),$$

where the second line follows from $u(\cdot)$ being strictly concave.[14] With actuarially fair prices, the households will buy full insurance and consumption does not depend on ω^t. Full insurance is, of course, only feasible in equilibrium if total resources do not vary across the different values of ω^t so in general, prices have to adjust to reflect not just probabilities, but also relative scarcity, across states.

Risk sharing For ω^t, take the ratio of equation (7.20) for household i and household j:

$$\frac{u'(c_{i,t}(\omega^t))}{u'(c_{j,t}(\omega^t))} = \frac{\lambda_i}{\lambda_j}.$$

Now solve for the consumption of household i in terms of that of household j

$$c_{i,t}(\omega^t) = u'^{-1}\left(\frac{\lambda_i}{\lambda_j} u'(c_{j,t}(\omega^t))\right).$$

Goods market clearing requires $\sum_i c_{i,t}(\omega^t) = \sum_i y_{i,t}(\omega^t)$, so we obtain

$$\sum_i u'^{-1}\left(\frac{\lambda_i}{\lambda_j} u'(c_{j,t}(\omega^t))\right) = \sum_i y_{i,t}(\omega^t). \tag{7.21}$$

Equation (7.21) relates the consumption of household j to the aggregate supply of goods and the Lagrange multipliers of all the households. Importantly, those Lagrange multipliers are constant across time so $c_{j,t}(\omega^t)$ varies over time as a function of aggregate income not as a function of $y_{j,t}$. This is a very important result for complete markets models: all idiosyncratic risk is insured away and consumption fluctuations only reflect aggregate risks.

[14] In a static context, an actuarially fair gamble is one in which the expected payoff is equal to the cost of the gamble. To adapt this definition to a dynamic context, let's say prices are actuarially fair if the price of any payoff at date t is equal to the price of any other payoff at date t that has the same expected payoff. The condition $p_t(\omega^t) = \bar{p}_t \times \pi_t(\omega^t)$ imposes this definition. To see this, consider a set of payoffs at different histories that can arise at date t given by $x_t(\omega^t)$. The date-0 value of such a portfolio of claims is $\sum_{\omega^t} x_t(\omega^t) p_t(\omega^t) = \bar{p}_t \mathbb{E}[x_t(\omega^t)]$, which only depends on the date and the expected payoff.

Aggregate and idiosyncratic risks Aggregate risks lead to movements in aggregate variables such as aggregate income or aggregate consumption while idiosyncratic risks affect an individual's circumstances but do not affect the aggregate. One person becoming unemployed is an example of an idiosyncratic shock while an event that changes the unemployment rate is an example of an aggregate shock. If the set of consumers, $\mathcal{I}$, is finite then an individual shock, by definition, is an aggregate shock, albeit a small one if $\mathcal{I}$ has many elements. If $\mathcal{I} = [0, 1]$, then the set is (even uncountably) infinite and one individual's shock is truly idiosyncratic, unless it is synchronized/correlated with the shocks of others. To illustrate, an interesting case is precisely that where the shock is "employment" (say, with income y_e for the individual) or "unemployment" (with income $y_u < y_e$). The date-t shock ω_t would then specify a whole function taking, for each $i \in [0, 1]$, the value e or u. The whole event tree would even be hard to imagine. A special case is where individuals' employment outcomes are independent draws with probabilities π_e and $\pi_u = 1 - \pi_e$, respectively. Then if we can appeal to a law of large numbers there would be no aggregate uncertainty: aggregate resources would be a deterministic value $\pi_e y_e + \pi_u y_u$. However, each individual faces uncertainty, though in this case markets can allow full insurance. This kind of model, where the law of large numbers is assumed to hold, is often used in macroeconomics.[15] Now imagine that π_e is random: an aggregate shock, which itself could, e.g., take on two values (say, high or low) as well vary over time. Then individuals' shocks are correlated, though if one conditions on the aggregate shock (high or low unemployment), their shocks can be thought of as purely idiosyncratic and uncorrelated. Our notation involving $\mathcal{I}$ and ω^t is abstract and meant to capture all these possibilities.

Sequential trading We can implement a complete set of markets with an alternative trading arrangement in which agents only trade securities that pay off in the next period and then trade again every period. This parallels our two ways to define equilibrium in deterministic contexts in Sections 5.2–5.3: we now merely have stochastic sequences.

For each event ω_{t+1} that can occur at $t + 1$, there is an asset traded at t that pays one unit at $t + 1$ if that event occurs and zero otherwise. These are known as Arrow securities. Let $q_t(\omega_{t+1}|\omega^t)$ be the price at t of a unit of consumption at $t + 1$ if event ω_{t+1} occurs. This price can depend on the history leading up to date t, ω^t, and is denominated in terms of consumption after history ω^t. Let $a_{i,t+1}(\omega^{t+1})$ be the amount of this asset held by household i. The budget constraint of the household is then

$$c_{i,t}(\omega^t) + \sum_{\omega_{t+1}} q_t(\omega_{t+1}|\omega^t) a_{i,t+1}(\omega^{t+1}) \leq y_{i,t}(\omega^t) + a_{i,t}(\omega^t). \tag{7.22}$$

Financial wealth, $a_{i,t}(\omega^t)$, becomes a state variable for the household's problem. This wealth allows the household to consume more than its income stream in the current period and future periods. When financial wealth is negative, the household must consume less than its income either now or in the future.

While agents only trade assets that pay off one period in the future, they can use these asset prices to value payoffs further in the future. One unit of goods at $t + 2$ after history ω^{t+2} has a value in date t of $q_{t+1}(\omega_{t+2}|\omega^{t+1}) \times q_t(\omega_{t+1}|\omega^t)$. In this product, the first term

[15] Such an assumption involves mathematical subtleties; see, e.g., Uhlig (1996).

discounts the unit of goods back to $t+1$ and the second term discounts it from $t+1$ to t. In general, we can define these discounts recursively as

$$\tilde{q}^t_{\tau+1}(\omega^{\tau+1}) = q_\tau(\omega_{\tau+1}|\omega^\tau)\tilde{q}^t_\tau(\omega^\tau)$$

with $\tilde{q}^t_t(\omega^t) = 1$. While the households do not trade assets for dates $\tau > t+1$ at date t, they do correctly anticipate the prices that will prevail in the future and the $\tilde{q}$ terms reflect these expectations.

At date 0, we assume that households have no financial wealth positive or negative because we assume that no trades have occurred prior to date 0 and so no household has a financial claim on any other household. Similar to models without uncertainty, the no Ponzi game constraint requires that the household could repay if it consumes nothing forever:

$$a_{i,t}(\omega^t) \geq -\sum_{\tau=t}^{\infty}\sum_{\omega^\tau}\tilde{q}^t_\tau(\omega^\tau)y_{i,\tau}(\omega^\tau). \tag{7.23}$$

Notice that this constraint rules out Ponzi games: it is the "natural borrowing limit," discussed in Section 4.3.1, now applying state by state.

We can now write the household's problem as

$$\max_{\{c_{i,t}(\omega^t),a_{i,t}(\omega_t)\}_{\forall t,\omega^t}} \sum_t\sum_{\omega^t}\beta^t\pi_t(\omega^t)u(c_{i,t}(\omega^t))$$

such that (7.22) and (7.23) hold for all t and ω^t.

A competitive equilibrium is a consumption allocation $c_{i,t}(\omega^t)$ for all i, t, and ω^t; asset positions $a_{i,t}(\omega^t)$ for all i, t, and ω^t; a price system $q_t(\omega_{t+1}|\omega^t)$ for all t, ω_{t+1} and ω^t such that (i) for all i, the consumption-savings plan is optimal taking the prices, borrowing constraints, and $a_{i,0} = 0$ as given; (ii) for all t and ω^t, the goods markets clear: $\sum_i (c_{i,t}(\omega^t) - y_{i,t}(\omega^t)) = 0$; (iii) for all t and ω^t, the asset market clears $\sum_i a_{i,t}(\omega^t) = 0$.

Definition 14 *A **sequential competitive equilibrium** is a set of stochastic sequences* $\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}$ *and* $\left\{a^\star_{i,t+1}(\omega^t) : \forall t, \omega^t\right\}$ *for each* $i \in \mathcal{I}$*, and* $\{q_t(\omega^t) : \forall t, \omega^t\}$ *such that*

1. *for each* i, $\left(\left\{c^\star_{i,t}(\omega^t) : \forall t, \omega^t\right\}, \left\{a^\star_{i,t+1}(\omega^t) : \forall t, \omega^t\right\}\right)$ *solves*

$$\max_{\{c_t(\omega^t):\forall t,\omega^t\},\{a_{t+1}(\omega^t):\forall t,\omega^t\}} \sum_{t=0}^{\infty}\sum_{\omega^t\in\Omega^t}\beta^t\pi(\omega^t)u(c_t(\omega^t))$$

subject to

$$c_{i,t}(\omega^t) + \sum_{\omega_{t+1}} q_t(\omega_{t+1}|\omega^t)a_{i,t+1}(\omega^{t+1}) = y_{i,t}(\omega^t) + a_{i,t}(\omega^t) \tag{7.24}$$

and (nPg)

$$a_{i,t}(\omega^t) \geq -\sum_{\tau=t}^{\infty}\sum_{\omega^\tau}\tilde{q}^t_\tau(\omega^\tau)y_{i,\tau}(\omega^\tau) \tag{7.25}$$

2. $\sum_i \left(c^\star_{i,t}(\omega^t) - y_{i,t}(\omega^t)\right) = 0$ *and* $\sum_i a^\star_{i,t+1}(\omega^t) = 0$ *for all* t *and* ω^t.

Here, requirement *2* has two conditions (for each date and state); one of these implies the other.

As in the case of certainty, the equilibrium allocation under sequential trading is the same as the one that arises with date-0 trading. To prove this, the key step is to add the stochastic sequence of budget constraints, multiplied by the appropriate prices, to arrive at a time-zero consolidated constraint (after using the nPg condition). Then, after seeing how the prices in the two settings map into each other, it becomes clear that the consumers solve the same problems in the two equilibrium definitions.

Spanning and complete markets Arrow securities are a convenient modeling device, but they are not recognizable as assets that we normally trade. Most real-world assets pay off in more than one state of nature. A system of markets would still be complete if we can construct portfolios of the available assets that have payoffs equivalent to a full set of Arrow securities.

Suppose there are S states of the world that might be realized at date $t+1$ and there are N assets traded at each date. We can construct the $S \times N$ payoff matrix D that lists what each asset pays in each state. A portfolio is a vector $\theta \in \mathbb{R}^N$ that lists the weights on each asset. The $S \times 1$ vector listing the payoff of a portfolio θ in each state of the world is given by $D\theta$.

The range of the matrix D is the space of all possible payoffs that can be constructed by making portfolios of the N assets. Let

$$\mathcal{M} \equiv \left\{ z \in \mathbb{R}^S : z = D\theta \text{ for some } \theta \in \mathbb{R}^N \right\}.$$

If $\mathcal{M} = \mathbb{R}^S$, the system of markets is complete. In this sense, a complete market means that one can construct a portfolio with any conceivable payoff vector. A system of markets will be complete if and only if $\text{rank}(D) = S$. If this rank condition is satisfied by the N assets, we say the assets **span the payoff space**. If a system of markets is complete, there are portfolios $\{\theta^{\text{A}}_j\}_{j=1}^S$ such that $D\theta^{\text{A}}_j$ pays one unit if state j occurs and zero units otherwise. The portfolios $\{\theta^{\text{A}}_j\}_{j=1}^S$ are the Arrow securities. For much more on this, see Chapter 16.

7.5 Competitive equilibrium in the growth model

We now use the framework of trade under uncertainty we just developed to define a competitive equilibrium for the stochastic growth model. We will state all of the assumptions here even though some of them were already introduced in Section 7.3.

We assume, for simplicity, that all individuals are identical. The representative household is endowed with k_0 units of capital and one unit of labor that is supplied inelastically. The intertemporal utility function is

$$\sum_t \sum_{\omega^t} \beta^t \pi_t(\omega^t) u(c_t(\omega^t)).$$

Output is produced according to

$$y_t(\omega^t) = A_t(\omega^t)F(k_t(\omega^{t-1}), 1),$$

where F has the usual neoclassical properties. The aggregate resource constraint is

$$k_{t+1}(\omega^t) + c_t(\omega^t) = (1-\delta)k_t(\omega^{t-1}) + y_t(\omega^t).$$

Turning to markets, the representative household accumulates capital and rents it to a representative firm in a spot market at price $r_t(\omega^t)$. The total, gross, return is $r_t(\omega^t) + 1 - \delta$, thus also inclusive of the undepreciated capital. Similarly, the household rents its labor to the firm in a spot market at a price $w_t(\omega^t)$.

The firm's problem is exactly as we have discussed before and results in the first-order conditions

$$r_t(\omega^t) = A_t(\omega^t)F_1(k_t(\omega^{t-1}), 1) \tag{7.26}$$

and

$$w_t(\omega^t) = A_t(\omega^t)F_2(k_t(\omega^{t-1}), 1). \tag{7.27}$$

The household's budget constraint is

$$c_t(\omega^t) + k_{t+1}(\omega^t) = (r_t(\omega^t) + 1 - \delta)k_t(\omega^{t-1}) + w_t(\omega^t).$$

We could allow the household to trade Arrow securities contingent on ω^{t+1} but, without another party to trade with, the representative household must have a zero position in each security in equilibrium. Therefore, although the consumer faces incomplete markets here, a full set of state-contingent assets would not change the equilibrium allocation. The household cannot hold a negative capital position, $k \geq 0$, but we will assume this constraint does not bind.

The Lagrangian of the household's problem is

$$\mathcal{L} = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \left\{ \beta^t \pi_t(\omega^t) u(c_t(\omega^t)) + \lambda_t(\omega^t) \left[(r_t(\omega^t) + 1 - \delta)k_t(\omega^{t-1}) + w_t(\omega^t) - c_t(\omega^t) - k_{t+1}(\omega^t) \right] \right\}.$$

Taking the first-order conditions with respect to $c_t(\omega_t)$ and $k_{t+1}(\omega^t)$ we have

$$\begin{aligned} \beta^t \pi_t(\omega^t) u'(c_t(\omega^t)) &= \lambda_t(\omega^t) \\ \lambda_t(\omega^t) &= \sum_{\omega^{t+1}|\omega^t} (r_{t+1}(\omega^{t+1}) + 1 - \delta)\lambda_{t+1}(\omega^{t+1}) \end{aligned}$$

and combining these we arrive at an Euler equation of

$$u'\left(c_t(\omega^t)\right) = \mathbb{E}_t\left[\beta u'\left(c_{t+1}(\omega^{t+1})\right)\left(r_{t+1}(\omega^{t+1}) + 1 - \delta\right)\right]. \tag{7.28}$$

A competitive equilibrium of this economy is a set of stochastic processes

$$\{r_t(\omega^t), w_t(\omega^t), c_t(\omega^t), k_{t+1}(\omega^t)\}_{\forall t, \omega^t}$$

such that $c_t(\omega^t)$ and $k_{t+1}(\omega^t)$ are optimal in the household's problem given the prices, the prices are set by competitive profit-maximizing firms in accordance with equations (7.26) and (7.27) and the resource constraint is satisfied.

Definition 15 *A **sequential competitive equilibrium** is a set of stochastic sequences* $\{c_t^\star(\omega^t) : \forall t, \omega^t\}$ *and* $\left\{k_{t+1}^\star(\omega^t) : \forall t, \omega^t\right\}$ *and* $\{(r_t(\omega^t), w_t(\omega^t) : \forall t, \omega^t\}$ *such that*

1. $\left(\left\{c_t^\star(\omega^t) : \forall t, \omega^t\right\}, \left\{k_{t+1}^\star(\omega^t) : \forall t, \omega^t\right\}\right)$ *solves*

$$\max_{\{c_t(\omega^t):\forall t,\omega^t\},\{k_{t+1}(\omega^t):\forall t,\omega^t\}} \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} \beta^t \pi(\omega^t) u(c_t(\omega^t))$$

subject to the nPg condition and

$$c_t(\omega^t) + k_{t+1}(\omega^{t+1}) = (r_t(\omega^t) + 1 - \delta)k_t(\omega^{t-1}) + w_t(\omega^t) \tag{7.29}$$

2. *for all* t *and* ω^t,

$$r_t(\omega^t) = A_t(\omega^t)F_1(k_t^\star(\omega^{t-1}), 1) \quad \text{and} \quad w_t(\omega^t) = A_t(\omega^t)F_2(k_t^\star(\omega^{t-1}), 1)$$

3. *for all* t *and* ω^t,

$$k_{t+1}^\star(\omega^t) + c_t^\star(\omega^t) = (1 - \delta)k_t^\star(\omega^{t-1}) + A_t(\omega^t)F(k_t^\star(\omega^{t-1}), 1).$$

In this definition of equilibrium the last condition is superfluous. This is, as in the deterministic case, because a constant returns to scale production function is homogeneous of degree one and by Euler's theorem we therefore have

$$r_t(\omega^t)k_t(\omega^{t-1}) + w_t(\omega^t) = A_t(\omega^t)F(k_t(\omega^{t-1}), 1) = Y_t(\omega^t).$$

If we substitute this into the household budget constraint we arrive at the aggregate resource constraint, which is the goods market clearing condition.

The competitive equilibrium allocation is the same as we obtain from the planner's problem. If we substitute equation (7.26) into (7.28) and we arrive at the same Euler equation as we obtained in the planner's problem, equation (7.13). Furthermore, if we substitute into (7.26) and (7.27) into the household budget constraint, we obtain the same resource constraint as applies to the planner's problem.

The equivalence between the planner's solution and the competitive equilibrium allocation is not affected by adding uncertainty to the model. Indeed, there is nothing fundamentally different about an economy with uncertainty as compared to a deterministic one. Instead of indexing goods just by time, we now index goods by time and histories. As a result, the first and second welfare theorems continue to apply.

In Section 7.3 we used the functional Euler equation to solve the planner's problem for the stochastic growth model. Due to the equivalence between the competitive equilibrium and the planner's problem, the solution we found is also the solution to the competitive equilibrium. In fact, in Appendix 7.A we formulate a recursive competitive equilibrium and derive the exact same functional Euler equation as we found for the planner's problem.

7.6 An incomplete-markets economy

In this chapter, we have mostly focused on competitive equilibria with complete markets. If a system of markets is incomplete, i.e., if some insurance contracts are not available, then the planner's problem and the competitive equilibrium will not coincide in general. In this case, the first welfare theorem fails to hold because some markets are missing (the markets for those insurance contracts). Here we will just give a brief introduction to incomplete-market models and return to this topic in later chapters.

Consider a consumer with preferences given by

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t).$$

The consumer receives a stochastic income stream y_t and can borrow or save in an asset that pays gross interest $1+r$, which is known and constant. Here we just describe the consumer's decision problem and we will take r as given.

Letting a_t be the agent's assets at the start of period t and $q = 1/(1+r)$, the budget constraint is

$$qa_{t+1} + c_t = a_t + y_t.$$

We assume there is some lower limit to how much the consumer can borrow, $a_{t+1} \geq \underline{a}$, where $\underline{a}$ could be the natural borrowing limit (the amount the consumer could repay if they received the lowest possible income realization in all future periods) or some more restrictive borrowing constraint.

If we assume the endowment process is a first-order Markov process, the consumer's problem can be stated recursively as

$$V(a, y) = \max_{a' \geq \underline{a}} \left\{ u(a + y - qa') + \beta \mathbb{E}\left[V(a', y') | y \right] \right\}.$$

The first-order condition of this problem is

$$u'(c)q = \beta \mathbb{E}\left[V_1(a', y') | y \right]$$

and the envelope condition is

$$V_1(a, y) = u'(c).$$

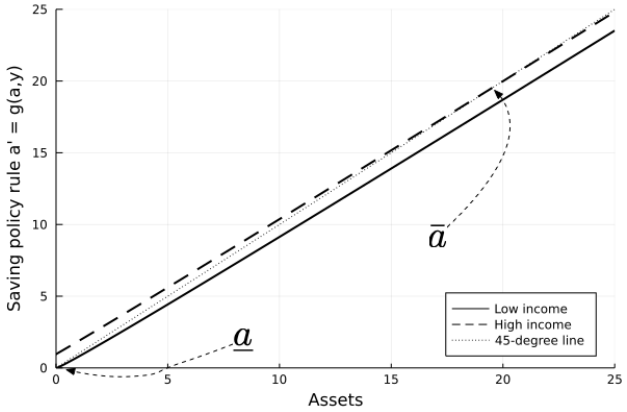
Combining these we have the Euler equation

$$u'(c) = (1+r)\beta \mathbb{E}\left[u'(c') | y \right].$$

This is an example of an incomplete-market environment because the single asset cannot insure the consumer against the income risk. In other terms, using the single asset with a constant return, the consumer can borrow and save but the payoff of their portfolio is independent of their income in the next period. As a result, their consumption will fluctuate in response to the realizations in their individual endowment. However, the consumers can partially self insure. They can accumulate savings that they can spend down if they receive low endowments in the future. In this way the consumer can partially smooth their

consumption. However, they will not in general choose to fully smooth their consumption. To see this, suppose a consumer wished to have a perfectly smooth consumption path. The only way they could do this is to set their consumption to a level that would be sustainable if they received the lowest possible income realizations at all future dates. If they choose a higher level, there would be a chance that they would be unlucky and have to reduce their consumption. But this extremely conservative plan will not be optimal even for a consumer with very high risk aversion. Instead, the consumer will choose to adjust their consumption level in response to the endowments they receive. In Chapter 11, and later in Chapter 21, we study problems of self insurance in much more detail.

(a) Decision rule for net saving $g(a, y) - a$



(b) Simulated stationary distribution of assets

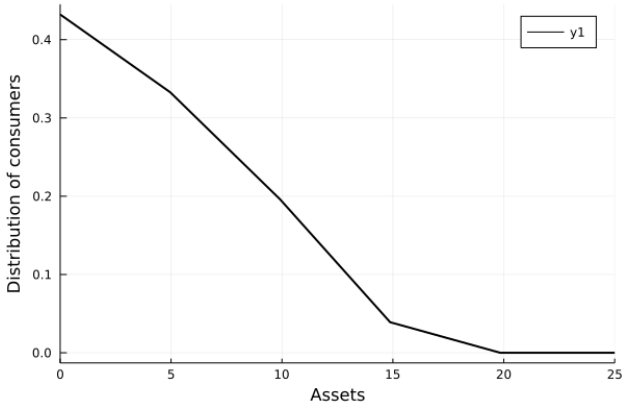


Figure 7.4: Decisions rules and stationary distribution for the incomplete-market consumption-savings problem.

Panel (a) of Figure 7.4 plots the decision rule $a' = g(a, y)$ for a version of this model in which income can take two values each period. When income is high, the consumer accumulates savings up to the point $\bar{a}$ where the upper line crosses the 45-degree line. When income is low, the consumer spends down their savings until it reaches the borrowing constraint. If the consumer begins with initial assets above $\bar{a}$, they will continuously spend down their assets regardless of their income until their assets reach $\bar{a}$. So in the long run the consumer will have asset holdings on the ergodic set $[\underline{a}, \bar{a}]$. Within this set, however, the consumer

will sometimes be moving towards higher asset levels (when they have high income) and will sometimes be moving towards lower levels (when they have low income). Contrast this saving rule with what we found in our deterministic steady state endowment economy in Section 5.4.1 where the decision rules were $a' = g(a)$ so they are simply the 45-degree line. In Figure 7.4, the decision rule of high-income consumers has a slope less than 45 degrees but an intercept above the 45 degree line while the decision rule of low-income consumers is on the 45-degree line at $\underline{a}$ but has a slope less than 45 degrees. The fact that the decision rules are above and below the 45-degree line is the source of partial insurance—those with high income put some of their "extra" resources into savings while those with low income draw down their savings.

Now suppose we simulate a large number of consumers each solving this decision problem and each receiving their own independent endowment process. The simulation produces a distribution of consumers over asset levels that reflects the fact that some consumers have been lucky and received many high endowments and some have been unlucky. Moreover, as consumption is a function of the consumer's assets and income, the simulation also produces a distribution of consumption levels. At a point in time, the distribution of asset holdings will depend in part on the distribution of asset holdings in the previous period. In the long run, the distribution will converge to a stationary distribution. Panel (b) of Figure 7.4 plots the stationary distribution of assets for our simulation.

In some contexts the complete market and representative agent assumptions are sensible simplifications of reality, but when they are not we often turn to incomplete-market models that build on the framework we have just presented. Such models can be used to study a wide range of issues and we will return to this topic in the chapters that follow.

Chapter 8

Empirical strategies and quantitative macroeconomics

8.1 Introduction

We often have questions that require quantitative answers. Can a certain theory explain the amount of growth or volatility we see in the data? How do the benefits of a proposed policy compare to the costs? Quantitative answers require empirical evidence. Perhaps there is evidence that directly speaks to question of interest. More often there is not and we need to bridge the gap between the available evidence and the answer to our question with theoretical models or assumptions.

The types of questions we contemplate typically require estimates of causal effects that measure the consequences of changing one aspect of the economy holding the other economic fundamentals constant. A key challenge for economists, and macroeconomists in particular, is that it is often impossible to perform experiments on the economy. Instead, we have to estimate these causal effects using data that reflect equilibrium outcomes in which many variables are jointly determined and responding to a variety of different shocks. In this chapter we will discuss four different strategies to address this challenge.

We start with two strategies that impose relatively fewer assumptions on the analysis starting with **natural experiments**. While we usually cannot perform experiments on the economy it may be the case that there is variation in the data that is effectively exogenous with respect to the other variables of interest. If we can find such natural experiments, then we can estimate causal relationships more or less directly. We then turn to techniques for identifying causal effects from time series data. A vector autoregression (VAR) is a statistical tool that expresses the dynamic evolution of a group of time series in terms of the lags of the same time series. In a **structural VAR**, we add additional identifying assumptions to a VAR to uncover patterns in the data that can be given a causal interpretation.

The last two strategies we consider rely closely on economic theory in the form of a fully-specified model. A model can be used as a laboratory in which we can perform experiments to measure causal effects and experiment with proposed policies. The key issue then is how we use data to inform the model and its parameters. **Structural estimation** treats the model as the data generating process and selects parameters to best explain the observed

data. The strategy of **calibration** considers each parameter of the model and seeks evidence that speaks to the value of that parameter. When we calibrate a model, we do not presume that the model explains all of the variation in the data we observe.

The rest of this chapter will demonstrate these four different strategies and how they would answer the question "how does GDP respond to an increase in government spending?" The answer to this question is often summarized by a single quantity known as the "fiscal multiplier." In section 8.2, we begin by setting out the core issue we face in identifying an exogenous change in government spending in order to estimate the fiscal multiplier. We then turn to the empirical strategies.

8.2 The identification challenge

The "fiscal multiplier" refers to the *causal* effect of an increase in government spending (G_t) on output (y_t). By causal effect we mean G_t increases independently of other forces affecting the economy. Suppose we have time series data on G_t and y_t. Then we could simply look at how strongly the two co-move. However, to estimate a causal effect rather than a correlation we need the increase in spending to be exogenous with respect to other variables of interest. For example, it could be the case that an increase in productivity creates more resources and the economy chooses to spend more on public goods. In that case, the increase in output is not caused by the increase in spending.

In order to illustrate the identification challenge, we will consider a simple static model of the economy. The economy is populated by a representative household with preferences

$$U_t = \log c_t - \frac{\ell_t^{1+\psi}}{1+\psi} + \gamma\eta_t \log G_t, \tag{8.1}$$

where c_t is consumption, ℓ_t is labor supply, G_t is spending on public goods. The parameter γ determines the steady state preference for public goods while η_t is an exogenous fiscal shock that shifts the value of public goods over time. For instance, government purchases may be especially useful in a time of war, which would correspond to an increase in η_t. The economy produces goods out of labor according to $y_t = A_t\ell_t$ where A_t is an exogenous TFP level. Goods can be used for consumption or public goods so we have $y_t = c_t + G_t$. Suppose a planner chooses $\{c_t, \ell_t, G_t\}$ to maximize the household's utility subject to $c_t + G_t = A_t\ell_t$. The solution to this problem results in the first-order conditions

$$\frac{A_t}{c_t} = \ell_t^{\psi}$$

and

$$G_t = \gamma\eta_t c_t.$$

Using the aggregate resource constraint to eliminate $c_t = y_t - G_t$ and the production function to eliminate $\ell_t = y_t/A_t$ we obtain

$$y_t = A_t\left(1 - \frac{G_t}{y_t}\right)^{-\frac{1}{1+\psi}} \tag{8.2}$$

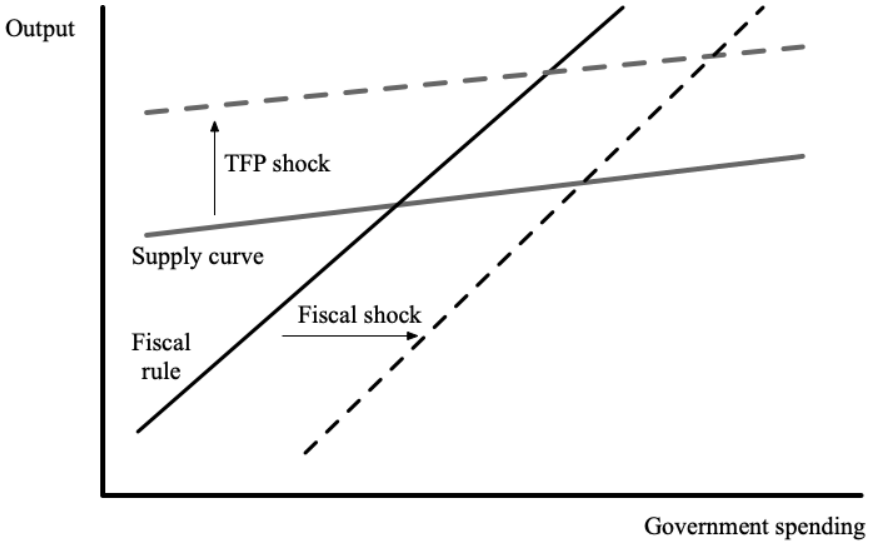


Figure 8.1: Equilibrium of supply curve (8.2) and fiscal rule (8.3).

and

$$G_t = \frac{\gamma \eta_t}{1 + \gamma \eta_t} y_t. \tag{8.3}$$

Equation (8.2) reflects the supply side of the economy. This equation depends on G/y due to wealth effects on labor supply. As more resources are devoted to the government and fewer to consumption, the marginal utility of consumption rises relative to the marginal disutility of work leading the planner to increase labor supply. Equation (8.3) is the fiscal policy rule—G_t increases if y_t or η_t increase. An increase in η_t clearly raises the benefit of public spending. An increase in y_t lowers the opportunity cost of allocating resources to G_t rather than c_t. Figure 8.1 plots these two relationships between G and y.

The fiscal multiplier is the change in output (in levels) divided by the change in spending (in levels). A fiscal multiplier equal to one means that one dollar more of spending raises GDP by one dollar. Crucially, the change in spending must be exogenous because the fiscal multiplier is meant to be a causal effect—it answers the question "what would happen if G increases *ceteris paribus*?" One way to calculate the fiscal multiplier in a model is to replace the spending policy rule with one in which G is simply exogenous. One would then consider the equilibrium response of output to exogenous changes in G. Graphically, the fiscal multiplier is the slope of the supply curve in Figure 8.1. An alternative approach to calculating the fiscal multiplier is to find an exogenous force that shifts the fiscal policy rule. In our model, the fiscal shock η_t is an exogenous shift of the fiscal policy rule. We can then calculate the fiscal multiplier as $(dy_t/d\eta_t)/(dG_t/d\eta_t)$. Or, in words, the change in y induced by a change in η scaled by the change in G that was induced by η.

Now suppose we have data on G_t and y_t over time. These data reflect movements in A_t as well as η_t. Changes in A_t shift the supply curve while changes in η_t shift the fiscal rule. As both of the curves in Figure 8.1 are subject to unobserved shocks, the empirical relationship between y_t and G_t does not reveal the slope of either curve. In order to calculate the fiscal multiplier, we need to isolate shifts in the fiscal rule that move the economy along the supply curve. The empirical strategies that follow all must confront this identification challenge.

Defining the fiscal multiplier in a dynamic economy

In a dynamic economy, a change in spending can be more or less persistent and the persistence of spending will matter for how the economy reacts. Moreover, an increase in spending at one date will lead to a change in output on impact and at subsequent dates. In this dynamic setting, the fiscal multiplier is not a single number, but depends on the details of the expected path of G_t and involves an impulse response of y_t.[a] The literature has used several methods to reduce this complexity to one dimension. Blanchard and Perotti (2002) define the multiplier as the ratio of the peak output response relative to the peak spending response. Alternatively, others follow Mountford and Uhlig (2009) in calculating the cumulative multiplier that integrates the impulse response of GDP and divides by the integral of the change in spending.

[a]Moreover, the response of y_t may depend on the state of the economy in which we start the experiment although in this chapter we will focus on first-order approximations to the dynamics of the economy that do not capture this state dependence.

8.3 Natural experiments

To assess the fiscal multiplier, we need a shift in fiscal policy that is unrelated to the state of the economy. Some of the largest changes in fiscal policy have occurred at times of war. It is reasonable to argue that these wars were not caused by economic conditions and so changes in government spending due to wars can be treated as exogenous. This is an example of using a natural experiment for identification. Given our knowledge of the economy, we may be able to find an exogenous variable (e.g. war spending) that is correlated with the endogenous variable of interest (e.g. total government spending). In our system (8.2)–(8.3), we seek a variable z_t that is correlated with G_t but uncorrelated with A_t.

A series of papers has used exactly this strategy of studying military spending to measure the fiscal multiplier. One challenge this strategy has to overcome is that the change in spending may have been known ahead of time. If wealth effects on consumption and labor supply are important mechanisms in the transmission of fiscal shocks to the rest of the economy, these effects should occur when the agents learn about the change in fiscal policy rather than when the spending actually occurs. For example, it may be clear that war is on the horizon before the government increases spending. Therefore the economic impact of the war spending may be felt before the spending occurs. To overcome this challenge, Ramey (2011) measures the change in the expected present value of defense spending each quarter by reading historical newspaper articles to understand the timing of information about military spending. This data is shown in Figure 8.2 where we see that there were large changes in expected military spending in WWII and the Korean War. Ramey estimates the impulse responses of government spending and GDP to this news and uses those impulse responses to estimate the fiscal multiplier. She finds a multiplier near 1 when the sample period includes WWII and multipliers near 0.7 when WWII is excluded.

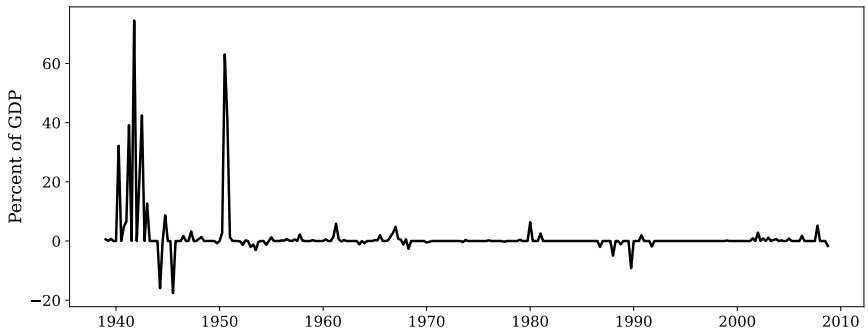


Figure 8.2: Change in the expected present value of military spending computed by Ramey (2011).

From natural experiments to impulse response functions: local projection

Let z_t be the measure of military spending news. As z_t is exogenous, we could simply regress output y_t on z_t. But perhaps there is a delay between the defense spending news and the output response. We could then regress y_{t+h} on z_t to estimate how output responds to spending news h periods after the news arrives. By varying h we can trace out a whole impulse response function. This procedure is called a local projection or Jordà (2005) projection.

Analyzing a natural experiment often requires using unconventional data sources. For example, Ramey used historical newspaper articles to construct her dataset. This is an example of the narrative approach in which one uses qualitative sources to obtain information that helps identify particular shocks or other developments in the economy. Romer and Romer (1989) is a well-known example within modern macroeconomics.[1] Their study sought to identify exogenous changes in monetary policy by reading transcripts of Federal Reserve meetings in order to find dates at which policy was changed despite the fact that policymakers felt that economic activity was at an appropriate level.

An alternative natural experiment strategy exploits the fact that military spending does not occur uniformly across an entire country but is concentrated in some regions. Using records of military purchases to understand the locations of military suppliers, Nakamura and Steinsson (2014) are able to construct a measure of military spending at the level of U.S. states. They show an increase in national military spending leads to a disproportionately large increase in spending in some states as compared to other states. A fiscal multiplier can then be estimated from the output movements in states with more spending relative to the output change in the states with less spending. The identifying assumption is that there are no other shocks which are both correlated with military spending in the time series and correlated with the fiscal loadings in the cross section. Nakamura and Steinsson estimate a multiplier of 1.5. The regional data provide useful evidence on the effects of government spending on the economy. It is worth noting, however, that a multiplier across states is not directly comparable to an economy-wide multiplier. When government spending increases at the national level, there may be changes in national taxes, interest rates, or commodity

[1] See also Romer and Romer (2023) for a review and update of this work.

prices. But these general equilibrium effects are common across states and therefore do not contribute to the estimated regional multiplier. For example, if the war spending is financed by national taxes, the tax impact will be the same (or similar) across states and therefore will not be part of the regional multiplier but it will be part of the national multiplier. Nakamura and Steinsson (2014) argue that the large regional multiplier they find is consistent with Keynesian theories in which an increase in demand leads to an economic expansion because it is the increase in demand that differs across states.

Natural experiments are often compelling because they isolate changes in the economy that are plausibly exogenous. Often times, this requires narrowing our focus to particular types of variation in the economy and we may worry that these events are not representative of the phenomenon as a whole. For example, it is possible that a general increase in government consumption has a different impact on the economy than military spending does. Moreover, there is no guarantee that we can find a convincing natural experiment for our question of interest. Therefore we may want to consider alternative empirical strategies.

8.4 Structural Vector Autoregressions

One option is to use the time series of aggregate government spending and ask how surprise innovations to this series affect the economy. The challenge here is that the surprise in spending may reflect an endogenous response of spending to some other development rather than a fiscal policy shock. The literature on structural vector autoregressions (VARs) seeks to overcome this challenge in order to uncover causal relationships in time series data.

Consider the process

$$B_0 Y_t = B_1 Y_{t-1} + B_2 Y_{t-2} + \cdots + B_J Y_{t-J} + \varepsilon_t, \tag{8.4}$$

where $Y_t \in \mathbb{R}^n$ is a vector of observed data, B_j is an $n \times n$ matrix of coefficients, and $\varepsilon_t \in \mathbb{R}^n$ is an i.i.d. stochastic innovation. Each equation in this system expresses a relationship between the current Y_t and lags of Y_t plus an error term. In a **structural vector autoregression** we give the equations and shocks a causal interpretation. We would like to think of one equation shifting without affecting other equations. For example, a TFP shock only shifts the production function without affecting other equations in the model.[2] We assume the structural shocks are uncorrelated with one another. It is without loss of generality to assume that the covariance matrix of ε_t is the identity matrix as we can rescale the B_j's by the standard deviations of the shocks.

To make things concrete, consider a log-linearized version of (8.2) and (8.3), assuming that the average value of η_t is 1,

$$\hat{y}_t = \frac{\chi}{1+\chi}\hat{G}_t + \frac{1}{1+\chi}\hat{A}_t \tag{8.5}$$

[2]In this context, the word "structural" is used in the manner it used in the study of simultaneous equation models in econometrics. In the structural system, the equations and shocks are given a causal interpretation. In other contexts, the term "structural" has a different meaning and refers, instead, to parametric models that are explicitly derived from economic theory.

and

$$\hat{G}_t = \hat{y}_t + \frac{1}{1+\gamma}\hat{\eta}_t, \tag{8.6}$$

where

$$\chi \equiv \frac{1}{1+\psi}\frac{\bar{G}/\bar{y}}{1-\bar{G}/\bar{y}}.$$

The shock in each equation has a clear interpretation: $\hat{A}_t$ is a shock to the supply-side of the economy while $\hat{\eta}_t$ is a shock to fiscal policy. The equations in the system cannot be estimated directly due to endogeneity concerns. In the first equation, G_t is correlated with A_t because it is determined by the solution to the whole system, which depends on A_t. But if we can obtain consistent estimates of the parameters, we would have most of the information we need to calculate the fiscal multiplier because it tells us the causal effect of $\log G$ on $\log y$.[3]

While we cannot estimate the structural system (8.4) directly we can transform it by premultiplying by B_0^{-1} to obtain the **reduced-form** VAR

$$Y_t = \underbrace{B_0^{-1}B_1}_{A_1} Y_{t-1} + \underbrace{B_0^{-1}B_2}_{A_2} Y_{t-2} + \cdots + \underbrace{B_0^{-1}B_J}_{A_J} Y_{t-J} + \underbrace{B_0^{-1}\varepsilon_t}_{u_t}. \tag{8.7}$$

As we have now eliminated the contemporaneous variables from each equation, we can estimate the equations of (8.7) via OLS, but the shocks we recover are not the structural shocks ε. Instead, the reduced-form residuals u_t are linear combinations of the structural shocks with weights given by B_0^{-1}. If our goal is to understand how a shift in fiscal policy affects the economy, we need to isolate a fiscal policy shock so the reduced-form VAR itself does not allow us to answer the question of interest.

As we assumed $\text{Var}(\varepsilon_t) = I$, it follows that $\text{Var}(u_t) = B_0^{-1}B_0^{-1\prime}$. In this way, the estimated covariance matrix of u_t gives us information about B_0. However, it does not give us all the information we need. Let F be some orthogonal matrix, which means $F'F = I$. Then for any candidate solution b such that $bb' = \text{Var}(u_t)$ there is an alternative solution bF' that also satisfies $bF'Fb' = \text{Var}(u_t)$. To resolve this problem, we need more information that allows us to eliminate all these rotations F except for one. This information does not come from the estimated reduced-form VAR but from identifying assumptions.

The matrix $\text{Var}(u_t)$ has dimension $n \times n$ but it is symmetric so it only contains $n(n+1)/2$ distinct values. To determine all of B_0 we need an additional $n^2 - n(n+1)/2$ restrictions on B_0.[4] These restrictions are identifying assumptions that must be motivated by theory or other data besides that contained in Y_t.

In our example system (8.5)–(8.6), we have $n = 2$ so we need one restriction on B_0. Our theoretical model implies that the coefficient on $\hat{y}_t$ in (8.6) is equal to 1 and we could use that as our identifying assumption. How does restricting a coefficient to 1 solve the identification problem? By defining $\hat{g}_t \equiv \hat{G}_t - \hat{y}_t$, we could rewrite (8.5)–(8.6) as

$$\hat{y}_t = \chi\hat{g}_t + \hat{A}_t \tag{8.8}$$

[3]The system (8.5)–(8.6) is written in logs while the fiscal multiplier is defined in levels. We can convert from logs to levels by multiplying the estimated effect in logs by $\bar{y}/\bar{G}$. Appendix 8.A.2 verifies this in the context of the model here.

[4]We are simply counting equations and unknowns here: the unrestricted B_0 has n^2 unknowns but due to symmetry $B_0^{-1}B_0^{-1\prime} = \text{Var}(u_t)$ only provides $n(n+1)/2$ equations. We therefore need to impose additional equations ("restrictions") to determine B_0.

and

$$\hat{g}_t = \frac{1}{1+\gamma}\hat{\eta}_t. \tag{8.9}$$

While G_t is endogenous, equation (8.9) makes clear that our theory asserts that g_t is exogenous and we can then attribute any movements in g_t to fiscal shocks. There is then no endogeneity problem in equation (8.8) because g_t is exogenous.

Recursive identification

The system (8.8)–(8.9) is an example of **recursive** identification. There is one equation that has no contemporaneous variables on the right-hand side, which allows us to identify shocks to the variable on the left-hand side of that equation. We can then treat that variable as exogenous in other equations and identify shocks to equations that only have exogenous variables on the right-hand side. In a larger system we could potentially identify all the shocks in a series of steps.

A recursive identification scheme is also called "Cholesky" or "triangular" identification. Recall that a Cholesky decomposition of a real positive definite matrix M gives a lower-triangular matrix L such that $LL' = M$. If we order the variables in Y_t so that the first variable does not depend on any other contemporaneous variable, the second variable may depend on the first but no other contemporaneous variables and so on, then a Cholesky decomposition of the covariance matrix VAR(u_t) gives B_0^{-1}.

Equation (8.9) imposes the restriction that $\hat{g}_t$ does not depend on $\hat{y}_t$. In an application with lagged variables, we could allow $\hat{g}_t$ to depend on lags of $\hat{y}_t$. So the identifying assumption only restricts the immediate impact of the shocks and not the dynamic relationships between the variables. This is the appeal of structural VARs, we can allow the dynamics of the economy to be quite flexible.

When we impose that certain elements of B_0 are equal to zero, we call those "timing restrictions" because we are imposing that certain variables do not depend on others within the same period but they can depend on those other variables with a delay. Blanchard and Perotti (2002) use timing restrictions to identify fiscal policy shocks in a structural VAR. They assume that fiscal policymakers cannot immediately (within a quarter) observe current GDP and adjust fiscal policy because data on GDP only arrives after the quarter is complete and because it takes time to pass legislation to adjust policy. This argument motivates a specification in which the contemporaneous response of spending to GDP is restricted to zero. Blanchard and Perotti's results are somewhat sensitive to the details of the specification but the multipliers they find are generally around one.[5]

Timing restrictions are not the only route to identification in a structural VAR. One approach that is widely used now is to use a natural experiment as an instrumental variable to identify a shock of interest in the VAR. In another approach, Uhlig (2005) proposed imposing restrictions on the signs of the responses of some variables following the shock of interest. Alternatively, Blanchard and Quah (1989) used the identifying assumption that only

[5]Recall that Blanchard and Perotti define the multiplier as the ratio of the peak output response to the peak spending response.

supply shocks can affect GDP in the long-run while demand shocks can lead to temporary fluctuations—a so-called "long-run restriction."

From natural experiments to impulse response functions using VARs

Let z_t be an exogenous variable, say a measure of military spending news. One way of computing impulse responses to shocks to z_t is to include z_t as an exogenous variable in a VAR. Specifically, one could use a recursive identification scheme with the assumption that z_t does not respond to any of the other variables in the system (i.e. it is ordered first). In her study of military spending, Ramey (2011) estimates the impulse responses by including the military news as an exogenous series in a VAR. This procedure is an alternative to a local projection

A limitation of structural VAR modeling is the invertibility assumption that the reduced-form innovations at date t are linear combinations of the structural shocks at date t. This assumption is implausible because it asserts that the variables in the VAR and their lags summarize all of the relevant information about the state of the economy and the only reason the VAR has a forecast surprise at date t is because there is a structural shock at date t. In contrast, suppose the estimated VAR system omits some useful information about the state of the economy. As this information is ignored, there will be a forecast surprise at t that is not due to a structural shock at t, but rather a symptom of the imperfect information at $t-1$. In general, it is not possible to test invertibility—you cannot be sure there is no other useful information. Instead, one can attempt to include a rich set of variables that plausibly capture the main factors relevant to forecasting. On the other hand, the number of parameters to estimate grows rapidly as the system expands leading to overfitting concerns. Resolving this tension between a richer information set and overfitting has been, and continues to be, an important issue for research.

The appeal of structural VARs is that they impose relatively few assumptions on the dynamics of the economy. On the other hand, there are also benefits that come with making more structural assumptions in the form of a model built on microeconomic foundations. With a full description of the economic environment, we can be more specific about economic mechanisms, the welfare consequences of shocks and policy decisions, and we can connect data from a variety of sources including micro-level data.

8.5 VARs and local projections

We often want to compute an impulse response function (IRF) that summarizes how one variable responds dynamically to innovations in another or in itself. Two common ways of estimating IRFs are local projections and VARs. This section describes the relationship between these estimation techniques, the IRFs that they seek to estimate, and how the two methods compare.

Let y_t be an $n \times 1$ vector of variables at date t. Without loss of generality, we can suppose that we are interested in the response of y_t to a change in the variable in the first position of the vector, $y_{1,t}$. Before going further, it is worth briefly mentioning identification. Nothing

we will discuss in this section has to do with a structural interpretation of innovations to $y_{1,t}$. If you have identified quasi-random variation in an instrument z_t you can simply include it as $y_{1,t}$ and use either method to estimate how outcomes of interest react to the instrument. Similarly, if you are interested in simply estimating dynamic relationships between variables without a structural interpretation you may again use either method.

A local projection constructs an IRF one horizon at a time. Letting $h \geq 0$ be the horizon, we consider the regression

$$y_{i,t+h} = \beta_{i,h} y_{1,t} + \sum_{j=1}^{J} A_{i,j} y_{t-j} + \varepsilon_{i,h}.$$

We run this regression for $h = 0, 1, \cdots, H$ and then the impulse response is $\{\beta_{i,h}\}_{h=0}^{H}$.

A (reduced-form) VAR specifies

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_J y_{t-J} + u_t,$$

where the A's are $n \times n$ coefficient matrices and u_t is an $n \times 1$ vector of innovations. Each row of this system is a regression equation and we can estimate the A's one equation at a time using OLS. When $J = 1$, the IRFs are easy to calculate. The impulse responses at horizon h are given by A_1^h, where the (i, j) element gives the response of $y_{i,t+h}$ to $\varepsilon_{j,t}$. Notice that at horizon 0, $A_1^h = I$ so the reduced-form shocks only perturb a single variable on impact. This is also true when $J > 1$, but now the dynamic IRFs are more complicated. In those cases, we can construct the companion form that expresses the system as a first-order system in an expanded set of variables. For example for $J = 2$ the companion form is

$$\mathcal{A} = \begin{pmatrix} A_1 & A_2 \\ I & 0 \end{pmatrix}$$

with

$$\begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix} = \mathcal{A} \begin{pmatrix} y_{t-1} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \end{pmatrix}.$$

The upper-left $n \times n$ block of $\mathcal{A}^h$ then gives the IRFs at horizon h. As before, the (i, j) element gives the response of $y_{i,t+h}$ to $u_{j,t}$.

These two estimation techniques as described above are estimating different things. If we assume $\{y_t\}$ is Gaussian (so least squares equates with a conditional expectation), the local projection will estimate

$$\mathbb{E}\left[y_{i,t+h} | y_{1,t} = 1, \{y_{t-j}\}_{j=1}^{J}\right] - \mathbb{E}\left[y_{i,t+h} | y_{1,t} = 0, \{y_{t-j}\}_{j=1}^{J}\right].$$

Notice that these expectations do not condition on $y_{j,t}$ for $j \geq 2$. In contrast, the (reduced-form) VAR assumes that there is no contemporaneous impact of one variable on another and is therefore estimating

$$\mathbb{E}\left[y_{i,t+h} | y_t = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}', \{y_{t-j}\}_{j=1}^{J}\right] - \mathbb{E}\left[y_{i,t+h} | y_t = \begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix}', \{y_{t-j}\}_{j=1}^{J}\right].$$

Plagborg-Møller and Wolf (2021) showed the IRFs from the reduced-form VAR can be combined to give the same impulse responses as the local projection.[6]

The local projection is very flexible in how it estimates impulse responses whereas the VAR imposes the assumptions of the statistical model to extrapolate the longer-horizon response from the short-term dynamics. With many lags, the VAR can be as flexible as the local projection, but in typical practice one has a moderate number of lags in a VAR.

The local projection is regressing y_{t+h} on $y_{1,t}$ (and lags). As h grows, there are many other shocks that have occurred in the meantime so the signal to noise ratio can be small leading to wide standard errors. The VAR, in contrast, regresses y_{t+1} on y_t (and lags). The VAR impulse responses are then extrapolated from these one-step-ahead dynamics. As the horizon h grows, the uncertainty around our estimate usually does grow at least for moderate h but less than for the local projection. So, the VAR will typically yield estimates with tighter confidence bands. However, if the auto-covariance function of the data is not well approximated by a VAR(J), the mis-specification of the statistical model can lead to larger bias in the VAR estimates than in the local projection estimates. In total, there can be a bias-variance tradeoff when we choose between a VAR and a local projection.

Figure 8.3 illustrates the bias-variance tradeoff. To generate the figure, we simulate a VAR process with two lags and then estimate local projections and a VAR with only one lag. We repeat the simulation 1,000 times and plot the 5th percentile, median, and 95th percentile of our estimated impulse responses. The dashed lines in the top panel show the true impulse responses from the VAR(2). In the bottom panel, we subtract the true IRF for ease of comparison. Clearly the local projection has more sampling variation but the VAR has more bias. For this example, the mean square errors are 0.0047 and 0.0027 for the local projection and VAR, respectively.

How to choose between the methods depends on your priorities. For point estimates, Li, Plagborg-Møller, and Wolf (2024) conducted a simulation study and found that for a researcher that wishes to estimate impulse responses with low mean square errors, the VAR approach is more appealing for many data generating processes. However, for inference, the bias in VAR estimates implies that standard confidence intervals typically do not cover the true impulse responses with the probability they are meant to (see Olea, Plagborg-Møller, Qian, and Wolf, 2024).

8.6 A model of fiscal policy

The final two empirical strategies we will discuss make use a of a structural model of the economy. A model can be viewed as an analytical tool that allows us to make sense of data and reach conclusions that are not immediately clear from the data alone. This is both an appeal of using models and a concern with using models. In some cases, the data do not

[6]The logic here is that we need to find the expected innovations to $y_{j,t}$ for $j \geq 2$ given a unit innovation to $y_{1,t}$. We then use these expected innovations in the other series as weights to combine the IRFs for $j = 1, \cdots, J$. Let Σ be the covariance matrix of the reduced-form VAR residuals. Let L be the lower-triangular Cholesky decomposition of Σ such that $LL' = \Sigma$. Now, consider the first column of L, call it $L_{\cdot,1}$ and normalize so that the first element of $L_{\cdot,1} = 1$. This vector gives the weights for a linear combination of the reduced-form VAR innovations that replicates the local projection estimand.

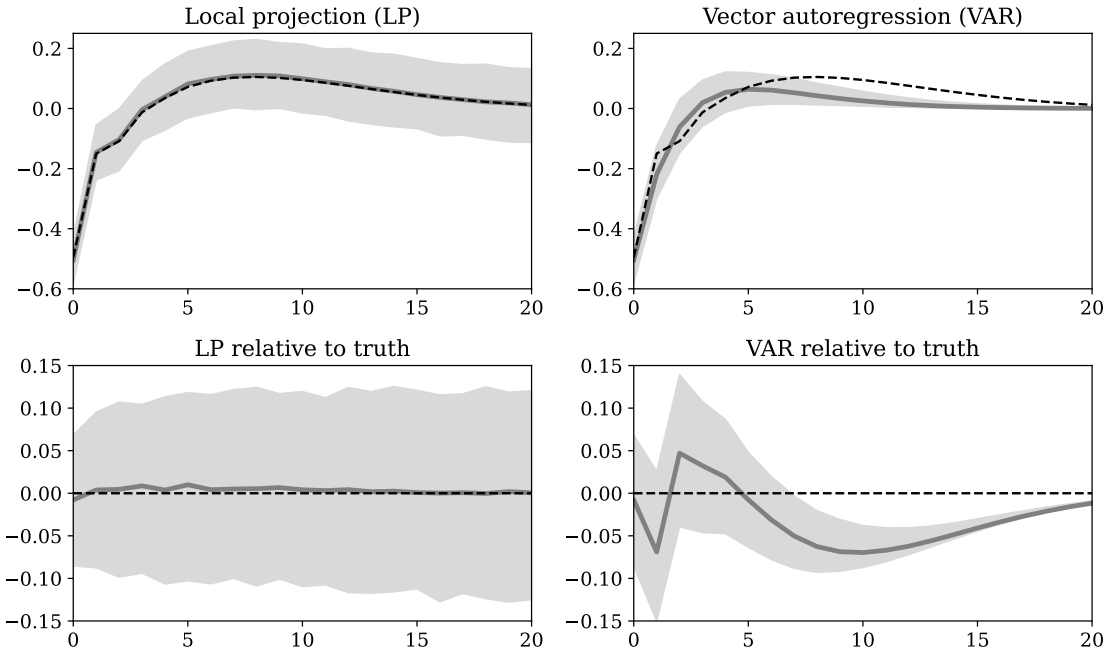


Figure 8.3: Bias-variance tradeoff in simulated data

Notes: We simulated 1,000 draws of length $T = 200$ from a VAR(2) in two variables. For each simulation draw, we estimate the impulse responses using local projection and a VAR but with only one lag. The solid lines are the median impulse responses at each horizon across the 1,000 draws. The shaded areas show the 5th and 95th percentiles at each horizon. The dashed lines show the true impulse responses. The lower panels show the differences between the estimated impulse responses and the true values.

speak directly to the question of interest because the type of experiment we are interested in has never occurred. We then have no choice but to use assumptions, typically in the form of a model, to extrapolate from the data we have observed to the question we are interested in. On the other hand, we may worry that the model assumptions drive our conclusions without a solid grounding in facts. To alleviate this concern, we need to use a model that allows for the economic channels that are most relevant to our question and allows the data to determine the relative strengths of these channels.

In our analysis of the fiscal multiplier, the static model presented in Section 8.2 only allows for one choice for the economy after an increase in government spending: the economy can work more or consume less. If we allow for investment, there is another possibility: the economy could reduce investment to free up resources for the government. In this section we will expand our model to incorporate this margin of adjustment. While the expanded model is a step in the right direction, it omits the distortionary effects of taxes and it omits Keynesian channels that are sometimes considered important in determining the effects of fiscal policy.[7]

8.6.1 The model

The main change we will now make relative to Section 8.2 is to include capital and investment. We will also now analyze a competitive equilibrium rather than a planner's solution because the model's implications for prices and payments to capital and labor are useful when calibrating the model.

A representative household has the preferences

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U_t,$$

where U_t is given in (8.1). Output is produced according to the production function

$$y_t = A_t k_t^{\alpha} \ell_t^{1-\alpha}. \tag{8.10}$$

Goods can be used for consumption by households, consumption by the government, or investment. Letting I_t denote investment, capital accumulates according to $k_{t+1} = (1 - \delta)k_t + I_t$ and the aggregate resource constraint is $y_t = c_t + I_t + G_t$. Combining the two we have

$$y_t = c_t + k_{t+1} - (1 - \delta)k_t + G_t. \tag{8.11}$$

For a given level of production, an increase in G_t will reduce the resources available for private consumption and investment.

We assume that households accumulate capital and rent it and their labor to firms in spot markets at rental rate r_t for capital and wage w_t for labor. Households must pay lump-sum taxes in amount T_t. The representative household's budget constraint is

$$c_t + k_{t+1} = (1 + r_t - \delta)K_t + w_t \ell_t - T_t.$$

[7] Chapter 15 discusses tax distortions and Chapter 18 will discusses new Keynesian models of the business cycle.

The government uses the tax revenue to finance its purchases so the government budget constraint is $T_t = G_t$. Imposing taxes on the household will affect their choices as they must consume less, save less, or work more in order to pay the taxes. However, as the taxes are imposed lump-sum, they do not affect the marginal returns to work and saving and therefore do not distort the household's choices away from the socially efficient choices.

There are two exogenous processes in this economy, which we assume follow AR(1) processes with unit mean:

$$A_t = (1 - \rho_A) + \rho_A A_{t-1} + \varepsilon_{A,t} \tag{8.12}$$

and

$$\eta_t = (1 - \rho_\eta) + \rho_\eta \eta_{t-1} + \varepsilon_{\eta,t}, \tag{8.13}$$

where $\varepsilon_{A,t}$ and $\varepsilon_{\eta,t}$ are i.i.d. with zero mean and variances σ_A^2 and σ_η^2, respectively.

The household's problem is to choose consumption, savings and labor supply to maximize its utility subject to its budget constraint taking the prices and taxes as given. The solution to this problem follows the same steps as in Section 7.5 so we will simply state the optimality conditions:

$$\frac{1}{c_t} = \beta \mathbb{E}_t \left[(1 + r_{t+1} - \delta) \frac{1}{c_{t+1}} \right] \tag{8.14}$$

and

$$\frac{w_t}{c_t} = \ell_t^\psi. \tag{8.15}$$

Equation (8.14) is the consumption Euler equation under uncertainty. Equation (8.15) is the first-order condition for labor supply that equates the utility of working an extra unit of time and consuming the income to the marginal disutility of working more.

The representative firm's problem is to maximize profits by hiring labor and capital and using these inputs to produce. As usual, the firm's first-order conditions equate marginal products and input prices:

$$r_t = \alpha \frac{y_t}{k_t} \tag{8.16}$$

and

$$w_t = (1 - \alpha) \frac{y_t}{\ell_t}. \tag{8.17}$$

We will assume that the government chooses taxes and spending to maximize the welfare of the representative household. In doing so, the government can anticipate that taxes cause households to change their behavior. Chapter 15 will discuss this type of interaction between the government and the private sector in more detail. Here we note that the solution to the government's problem satisfies the same first-order condition as the planner's problem, which equates the marginal benefit of spending to the marginal cost of the household consuming less

$$\frac{\gamma \eta_t}{G_t} = \frac{1}{c_t}. \tag{8.18}$$

An equilibrium of this economy is a set of stochastic processes $\{k_t, c_t, y_t, \ell_t, G_t, r_t, w_t, A_t, \eta_t\}$ that satisfy equations (8.10)-(8.18). In what follows we will discuss functional forms and parameter values for the model. Once those elements are in place, we can compute a solution

to the model via several methods. To analyze the dynamics of the economy, we will use a linear approximation around the deterministic steady state as discussed in Section 7.3.4.[8]

8.6.2 The fiscal multiplier

To develop intuition for the fiscal multiplier in this model, we can begin by supposing that investment does not respond to the change in spending. As investment is unchanged, there are only two ways to free up the resources for the extra spending: work more or consume less. If the economy simply consumes less, then output is unchanged and the fiscal multiplier will be zero. On the other hand, if the economy simply works more, then output increases one-for-one with spending and the multiplier will be one. From the perspective of households, an increase in government spending means an increase in taxes and therefore it induces a negative wealth effect that causes them to consume less and work more so the actual outcome is somewhere between these two extremes and the fiscal multiplier will be between zero and one.

If we allow investment to respond, there is a third way to pay for the government spending: invest less. If the economy invests less, it can free up resources for the government but at the cost of having a smaller capital stock in the future. With less capital in the future, output is lower. By reducing investment the households have extra resources that they can use to consume more and work less. To the extent they work less, the fiscal multiplier will be lower. If the change in government spending is short-lived, reducing investment will help to smooth consumption—the economy reduces the capital stock to finance the government and then builds the capital stock back up once the government spending returns to normal. However, if the spending shock is highly persistent, then this "borrow from the future" strategy is not as effective because the economy needs to finance the government now and in the future. Therefore the fiscal multiplier is increasing in the persistence of the shock to spending.

8.7 Structural estimation

Structural estimation treats the economic model as a statistical model of the data we observe. For different configurations of the structural parameters, the observed data are more or less likely to have occurred. Using likelihood methods (maximum likelihood or Bayesian estimation) we can then estimate the parameters of the model from the observed data. To start, we will continue with the simplified model from Section 8.2 before generalizing to the richer model from Section 8.6.

To estimate the model with likelihood methods we need to make assumptions about how the shocks are distributed. For the sake of discussion, let us suppose A_t and η_t are i.i.d. with densities p_A and p_η. Suppose further that our observable variables are time series of government spending and output from dates 1 to T. We then ask how likely these data were for a given set of parameters. For a given date t, we can use (8.2) and (8.3) to solve for the

[8] See Appendix 8.A.1 for how to calculate the steady state of this model.

values of A_t and η_t that are implied by the observed y_t and G_t

$$A_t = y_t \left(1 - \frac{G_t}{y_t}\right)^{\frac{1}{1+\psi}}$$

and

$$\eta_t = \frac{G_t/y_t}{\gamma(1 - G_t/y_t)}. \tag{8.19}$$

The likelihood of these observations is then $\text{Prob}(G_t, Y_t|\psi, \gamma) = p_A(A_t) \times p_\eta(\eta_t)$ and the likelihood of the full sample is then

$$\text{Prob}(\{G_t, Y_t\}_{t=1}^T |\psi, \gamma) = \prod_{t=1}^{T} \left(p_A(A_t) \times p_\eta(\eta_t)\right).$$

Maximum likelihood estimation treats the likelihood function above as the objective function to maximize. Bayesian estimation uses Bayes rule to derive a posterior probability distribution of the parameters conditional on the data. Doing so requires a prior distribution of our beliefs over the parameters before observing the data. Once we have estimated the parameters, we can differentiate (8.2) to find the implied fiscal multiplier (see Appendix 8.A.2 for a derivation of the fiscal multiplier in this model).

It is worth considering how structural estimation resolves the identification challenge discussed above. Equation (8.3) imposes a functional form on the fiscal rule that implies the G/y ratio is not affected by TFP shocks. Therefore, any changes in G/y are due to fiscal shocks. We are then able to infer the fiscal shocks from movements in G/y as shown in equation (8.19). A reasonable criticism of this approach is that the functional form assumption for the fiscal rule may be mis-specified, leading to mis-identification of the fiscal shocks.

Calculating the likelihood for the static model is straightforward because we can solve for η_t and A_t directly from the model equations. In actual applications this typically won't be possible in part because there are unobserved state variables. For example, with the production function $y_t = A_t k_t^\alpha \ell_t^{1-\alpha}$, if k_t and A_t are unobserved, then we cannot say if output per hour is high because A_t is high or because k_t is high. The process of using observed data to infer to the position of an unobserved state is known as filtering.

We will give a brief overview of how one can use filtering techniques to construct the likelihood for the expanded model with capital. The model has three state variables (k_t, A_t, and η_t). We do not observe A_t or η_t. While we have measures of the capital stock, they are likely contaminated by substantial measurement error so we often treat k_t as unobserved, too. We will suppose we can observe the time series of y_t and G_t.

The solution of the model provides a policy rule for investment in the form

$$k_{t+1} = f(k_t, A_t, \eta_t).$$

This rule, along with the laws of motion for the stochastic processes (8.12) & (8.13), define a stochastic difference equation for the state vector $X_t \equiv (k_t, A_t, \eta_t)'$

$$X_{t+1} = F(X_t, \varepsilon_{t+1})$$

where $\varepsilon_{t+1} \equiv (\varepsilon_{A,t+1}, \varepsilon_{\eta,t+1})'$ is a vector of shocks. Similarly, we can use the other equations of the model to express the observed data $Y_t \equiv (y_t, G_t)'$ in terms of the states and we can use $Y_t = H(X_t)$ to represent this relationship. The functions F and H are non-linear functions that depend on the model parameters. If we linearize these equations we obtain

$$\hat{X}_{t+1} = \mathcal{A}(\theta)\hat{X}_t + \mathcal{B}(\theta)\varepsilon_{t+1} \tag{8.20}$$

$$\hat{Y}_t = \mathcal{C}(\theta)\hat{X}_t, \tag{8.21}$$

where $\mathcal{A}(\theta)$, $\mathcal{B}(\theta)$, and $\mathcal{C}(\theta)$ are coefficient matrices that depend on the model parameters, which we collect in the vector θ, and $\hat{X}$ is the deviation of X from steady state.

It is common to assume that ε is a Gaussian random variable because this allows us to use a particularly convenient filtering algorithm: the Kalman filter. We will specifically assume $\varepsilon \sim N(0, I)$. The Kalman filter generates estimates of the state at date t give information through $t-1$. These estimates are uncertain and the linear-Gaussian framework implies $X_t \sim N(X_{t|t-1}, P_{t|t-1})$, where $X_{t|t-1}$ is the conditional mean of X_t given information through date $t-1$ and $P_{t|t-1}$ is the covariance matrix that reflects our uncertainty over X_t. Given (8.21), we then have a distribution over Y_t that is also Gaussian with a mean $\mathcal{C}(\theta)X_{t|t-1}$ and covariance matrix $\mathcal{C}(\theta)P_{t|t-1}\mathcal{C}(\theta)'$. This gives us the likelihood of Y_t given data through $t-1$. We then can form the likelihood of full sample by applying this same logic to all dates t:

$$\text{Prob}(\{Y_t\}_{t=1}^T|\theta) = \prod_{t=1}^{T} \mathcal{N}\left(Y_t \quad | \quad \mathcal{C}(\theta)X_{t|t-1} \quad , \quad \mathcal{C}(\theta)P_{t|t-1}\mathcal{C}(\theta)'\right),$$

where $\mathcal{N}(Y|\mu, \Sigma)$ is the density of multivariate normal distribution with mean μ and covariance Σ evaluated at Y.[9] Once we have constructed the likelihood we can proceed with maximum likelihood or Bayesian estimation of the parameter vector θ.

To review, we have described a widely-used method for structurally estimating macro models. For a given θ, one solves for a linear approximation to the dynamics of the economy. One then uses the Kalman filter to filter the observed data, estimate the unobserved state, and construct the likelihood. Use of the Kalman filter assumes the shocks to the model are Gaussian. There exist methods for handling non-linear and non-Gaussian models, but the tractability of the Kalman filter leads linear-Gaussian models to be used in many applications.

Structural estimation treats the model as the data generating process. We next turn to calibration, where the goal is not to fully explain the data, but rather the gauge the strength of the key mechanisms embedded in the model.

Limited-information estimation

In this section, we have focused on likelihood methods or "full-information" estimation because that is a very common approach to structural estimation within macroeconomics. However, one could also consider limited-information methods that use par-

[9] Notice that we require a belief about the initial distribution of X_1 given no data (i.e. $X_{1|0}$ and $P_{1|0}$). In stationary models, it is common to assume that the initial state is drawn from the system's invariant distribution.

ticular moments of the data as estimation targets. One might estimate a structural parameter by using a single equation from a larger model and using the generalized method of moments. Such an approach still requires identification to obtain consistent estimates of the parameters and a benefit of specifying the full model is that the argument for identification can be laid out in concrete terms. Alternatively, estimated impulse response functions can be used as estimation targets. This procedure, known as impulse response matching, can be a bridge between the structural VAR or natural experiment approaches and fully structural modeling.

8.8 Calibration

Suppose we have some data of interest. This could be simple moments we calculate such as the volatility of GDP or it could be the result of some other analysis such as a study of a natural experiment. We can then ask if the mechanisms embedded in the model are able to account for this data of interest. To answer that question, we need parameters for the model. One option is to use the data of interest in picking the parameters, but this is not so compelling as it only tells us that there exist some parameter that can explain the data. Instead, we might seek other sources of information about the parameters. Then, with the parameters already set, we compare the model to the data of interest. This is the approach taken in calibration. In calibrating a model, we must be clear about what mechanisms our model uses to explain the data of interest and then take care to choose the parameters of the model with information that speaks to the strength of the relevant mechanisms.

In our analysis of fiscal policy, the data of interest could be the empirical estimates discussed above, which put the fiscal multiplier near one. Section 8.6 highlighted several considerations that are important to the size of the fiscal multiplier. The multiplier will be larger if households are more willing to vary their labor supply than their consumption. The fiscal multiplier will also be larger if the economy is unwilling or unable to reduce investment. As we calibrate the model, we should pay special attention to the parameters that affect these channels as they ultimately will determine the fiscal multiplier.

Chapter 3 has already discussed a calibration strategy for some of the parameters in our model. The parameter α determines the share of aggregate income that accrues to capital, which we observed to be in vicinity of 1/3. Chapter 3 also uses the average ratio of investment to capital, which was found to be 0.076, to calibrate the depreciation rate. The model presented here does not have population growth or trend productivity growth, so the only reason to invest in steady state is to replace the depreciated capital. The 0.076 ratio is for annual investment (a flow variable e.g. the investment spending over the course of a year) relative to the value of the capital stock (a stock variable e.g. the value at the start of the year). For the analysis of short-run dynamics, we often specify a model time period to be one quarter of a year. In quarterly terms, the investment-capital ratio is $0.076/4 = 0.019$ and we will set the depreciation rate $\delta = 0.019$.[10] From the Euler equation, we can

[10]The true depreciation rate is lower as some of the observed investment is explained by a growing economy. One could then argue that we should set δ lower, but we chose not to do that. If the economy invests less to finance government spending, it will move away from its steady state capital-output ratio. In reality, the decline in the capital-output ratio will reflect both depreciation and a growing economy. We will capture

see that the steady state return to capital is determined by $1/\beta$. Using (8.14) and (8.16) we obtain $\beta = (1 + \alpha\bar{y}/\bar{k} - \delta)^{-1}$. Inserting the observed capital-output ratio[11] of 3.3×4 into this equation yields $\beta = 0.994$.

The preceding parameters are quite closely connected to empirical counterparts, which allows for a straightforward calibration strategy. The inverse elasticity of labor supply, ψ, is more difficult to calibrate. The Frisch elasticity of labor supply is the percentage change in labor supply following a percentage change in the wage holding the marginal utility of consumption constant. With the preference specification used in the model, the Frisch elasticity is exactly $1/\psi$. A natural way to calibrate this parameter is to look for data on transitory changes in wages and see how much labor supply reacts. Wage fluctuations can reflect changes in labor supply as well as changes in labor demand. To identify the slope of the supply curve, we need to isolate changes in wages that are not themselves a reflection of labor supply shocks. One line of literature uses changes in tax rates as a source of variation in after-tax wages. This literature tends to find relatively small labor supply elasticities on the order of $1/2$ so $\psi \approx 2$ (see Chetty, 2012).[12]

In our model, a key mechanism is that the government imposes a lump-sum tax on households that makes them poorer and may cause them to work more through a wealth effect. The parameter ψ also determines the strength of this wealth effect in addition to the Frisch elasticity. We therefore want to be sure that our choice of ψ implies a reasonable magnitude for this wealth effect on labor supply. The ideal evidence on wealth effects on labor supply would involve random variation in wealth unrelated to other developments in the economy. Changes in wealth due to aggregate events, like an economic boom, are not ideal because the change in the economy that led to the boom may have also affected other aspects of the economy such as wages. Several studies have looked at the labor supply of individuals who win lottery prizes. Golosov, Graber, Mogstad, and Novgorodsky (2021) use data on lottery winners and compare their labor supply before and after they win. The authors find that winning \$100 leads, on average, to a \$2.30 drop in annual earnings. What does this imply for ψ? Solving equation (8.15) for ℓ and differentiating with respect to c we obtain $\frac{d\ell}{dc} = -\frac{1}{\psi}\frac{\ell}{c}$. We can multiply both sides by $4w$ to obtain the change in annual earnings. If we assume households smooth their consumption perfectly, they will consume the annuity value of their extra wealth. So receiving \$100 causes them to increase their consumption by $r/(1+r) \times \$100$. Putting these pieces together we have

$$\text{change in annual earnings} = -\$2.30 = -\frac{4}{\psi}\frac{w\ell}{c}\frac{r}{1+r}\$100.$$

We will use a quarterly interest rate of $r = 1\%$. The ratio $w\ell/c$ is quarterly earnings over quarterly consumption. Among households that work, this ratio is somewhat above 1 due to the life-cyle savings profile and we will set it to 1.2. Solving for ψ, we find $\psi = 2.1$. This is reassuring in that a value of ψ near 2 is consistent with evidence on the Frisch elasticity and with evidence on wealth effects.

both forces here by having a higher depreciation rate.

[11] As discussed in Chapter 2, the value of the capital stock relative to annual GDP is about 3.3. We multiply by 4 to express the ratio in terms of quarterly output rather than annual output.

[12] Some take the view that aggregating across a population of households changes the effective labor supply elasticity for the representative household. We will discuss this possibility in Chapter 14.

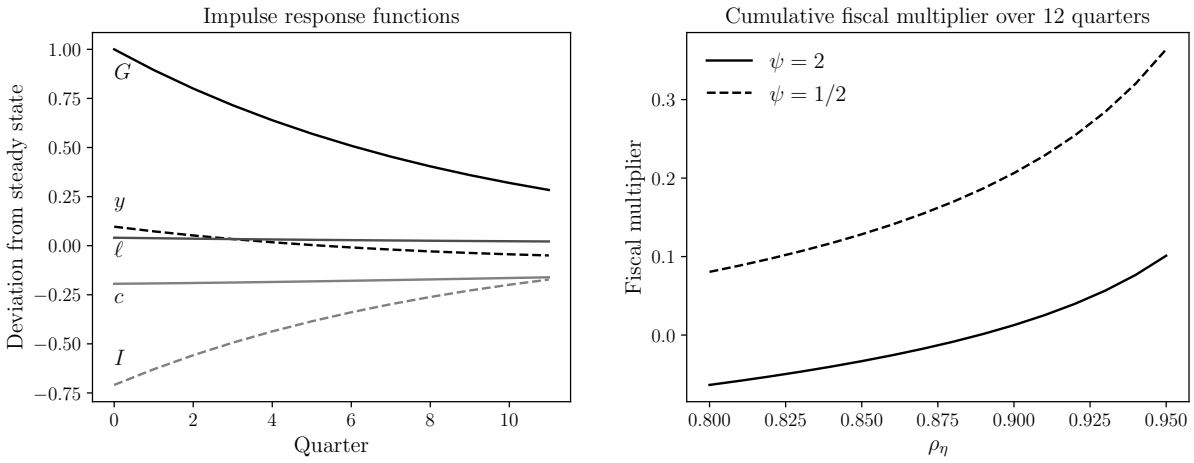


Figure 8.4: Impulse responses following a fiscal shock (left). Cumulative fiscal multiplier over the first 12 quarters (right).

In steady state, (8.18) becomes $\gamma = \bar{G}/\bar{c}$. In U.S. data from 1994 to 2023, the ratio of public consumption to private consumption has fluctuated in the range of 25% to 30%, while it was higher in the past (see Chapter 15 for data on the government's role in the economy). We will set $\gamma = 0.3$.

The model specifies exogenous stochastic processes for A_t and η_t. As our goal is to quantify the fiscal multiplier, we are interested in the economy's response to an exogenous change in fiscal policy, which we will generate with a shock to η_t. The presence of technology shocks will lead to consumption volatility and generate precautionary savings motives, which will have a small bearing on the fiscal multiplier. When we use a linear approximation of the model equations to analyze the economy, we implicitly assume these precautionary motives away. As a result, the economy's response to a fiscal shock will be the same regardless of whether or not technology shocks are included in the model. Additionally, the linear solution scales one-for-one with the size the shocks so the magnitude of the fiscal shock will not affect the fiscal multiplier. It follows that we do not need to specify ρ_A, σ_A^2 or σ_η^2. The fiscal multiplier does, however, depend on the persistence of the fiscal shock ρ_η. Our approach will be to report the fiscal multiplier for a range of values of ρ_η.

The left panel of Figure 8.4 shows the impulse responses to a fiscal shock (a positive realization of ε_η).[13] We have assumed that the shock is fairly persistent ($\rho_\eta = 0.9$) and G_t remains elevated for several years. In response to the shock, the economy works more, consumes less, and invests less. Quantitatively, we see that the bulk of the adjustment comes through investment and the change in labor supply is quite small. Initially, the increase in labor supply raises output. Over time, the decline in investment reduces the capital stock and, after 5 quarters, output falls below steady state. The cumulative fiscal multiplier—integrating the output and spending responses over the first 12 quarters—is close to zero.

In a sensitivity or robustness analysis we consider alternative calibrations and document how our conclusions are or are not affected by reasonable changes in the parameters. The right panel of Figure 8.4 shows that this conclusion is sensitive to the persistence of the

[13] We have scaled the shock so that G_t rises by one unit on impact, but this is just a normalization given the linear solution method.

spending shock. As ρ_η increases, reducing investment is less effective in smoothing consumption leading labor supply to adjust more resulting in a larger multiplier. Similarly, if ψ is smaller, so labor supply is more elastic, more of the adjustment occurs through labor supply and the fiscal multiplier is larger.

Model validation

Suppose that you have calibrated your model, but you still have additional empirical evidence that is relevant to your question. How can you incorporate this extra evidence into your analysis? One option is to ask whether your calibrated model can also match the additional facts. This step is called "model validation." The evidence employed in model validation is often related to the question of interest and perhaps less closely related to specific parameters. The point of model validation is to determine whether the model's answers can be trusted.

Earlier in the chapter we found fiscal multipliers closer to one while here we find multipliers closer to zero. From this analysis, it appears that wealth effects are unable to explain fiscal multipliers as large as those in the empirical estimates. We should note that this conclusion depends on the full model and it may be that wealth effects could be more powerful within a different model. Another possibility is that the model we consider here omits certain channels that make fiscal policy more powerful in stimulating output. The Keynesian view of fiscal policy is that it stimulates household incomes leading consumption to increase. This view contrasts quite strongly with the neoclassical view where government spending induces a negative wealth effect and a reduction in consumption.[14]

To conclude this section, we can draw some general principles for calibration. (i) *Understand the economics*: In calibrating a model, one has to make choices and the guiding principle is to obtain an accurate account of the strength of the main forces at play in the analysis. In order to make these choices, one must understand the economics of the problem. (ii) *Discipline the mechanisms*: Wherever possible, one should seek evidence that speaks directly to the strength of the mechanisms at the heart of the analysis. Microeconomic studies that cleanly isolate quasi-experimental variation in the economy are often an attractive source of information about the strength of mechanisms. (iii) *Robustness*: It is important to determine, and report, to what extent the quantitative results are sensitive to the parameter values chosen and to the exact way in which the model formalizes the mechanism in focus.

The conclusions one draws from a calibrated model are of course shaped in large part by the model itself. This is by design. Thus, the calibration example here asks about how large a fiscal multiplier would be if one only considered the mechanisms captured by the model at hand. The question is thus a partial one, to the extent one thinks the model lacks important elements. For example, if one thinks a Keynesian demand mechanism could be important then a next step is to add elements to the model, again by careful selection of parameters

[14] Chapter 18 discusses New Keynesian theories of the business cycle, which allow an increase in the demand for goods to increase output. Galí, López-Salido, and Vallés (2007) develop a New Keynesian model of fiscal policy and obtain fiscal multipliers near one.

relevant to the demand mechanism and evaluate again.

8.9 Taking stock

We have described some of the main methods that macroeconomists use for empirical and quantitative analysis. A reasonable question is "which one should I use?"

Macroeconomists increasingly value natural experiments as sources of identification. These natural experiments could occur at the national level, as in the military spending example above, or at a micro-economic level, as in the lottery winner example discussed in Section 8.8. In many cases, the natural experiment does not directly answer the question of interest, but it provides useful information about the economy that we can use as a calibration target for a model.

The strength of VAR analyses is that they place fewer restrictions on the dynamics of the economy than do full structural models. For example, in Figure 8.4 we see that in our structural model a government spending shock induces an immediate increase in output that then fades and a persistent decrease in consumption. These patterns are inherent to the model. If we calibrate or structurally estimate the model, the data will not be able to change these conclusions. In contrast, a VAR would allow more flexibility for the data to determine the shapes of the impulse responses.

Models take center stage when we lack data because they allow us to answer questions for which we do not have direct evidence. A model can serve as a bridge between the data we observe and the questions we want to answer. Our discussion of calibration is a good example—we can use data on capital, investment, and output from the national accounts along with data on how labor supply responds to taxes and lottery winnings to investigate a question about government spending. Models are also central when we are interested in understanding mechanisms at work in the economy.

It is worth noting that researchers often use multiple empirical methods in the same study. Natural experiments are often used to establish empirical results that may be explained by a structural model or may serve as a calibration target for a structural model. Regardless of which method you use, the success of your argument hinges on making a clear and compelling connection between your conclusion and the evidence that supports it.

Part II

Tools

Chapter 9

Continuous-time analytical techniques

In most chapters of this book, we develop models using a discrete-time framework. Discrete time is intuitive, aligns well with how we often approach data, and is typically more convenient for computation—especially when implementing numerical methods that require discretizing variables (see Chapter 10). However, in some cases, working in continuous time provides particular advantages. Certain economic problems become mathematically more tractable in continuous time, and key relationships—especially those involving differential equations—can be derived more transparently with "paper and pencil" techniques. Continuous-time models are especially helpful when we want to analyze economic variables that evolve smoothly over time, rather than in discrete jumps.

While discrete time remains the primary analytical tool throughout this textbook, the usefulness of continuous-time models will become clear in specific applications (such as those in Chapter 13). This chapter introduces the core analytical techniques specific to continuous-time models. Although most macroeconomic research and applications remain grounded in discrete time, it is important to become comfortable with continuous-time methods, as they offer valuable intuition and tools in particular research settings.

Discrete vs. continuous time: an intuitive overview Before introducing the formal notation and equations, it is helpful to clarify the distinction between discrete- and continuous-time models. In discrete time, we track economic variables—such as consumption, capital, or employment—at specific intervals (for example, monthly, quarterly, or yearly). This approach fits naturally with how economic data are typically measured and reported, and it facilitates computational work. Continuous-time models, in contrast, conceptualize time as a continuous flow, rather than a sequence of distinct intervals. Here, economic variables evolve smoothly at every instant, and their behavior is described by differential equations instead of difference equations.

To illustrate, consider the example of capital accumulation in an economy. In a discrete-time model, we observe the capital stock at the end of each period and model its change from one period to the next. With a continuous-time model, we track the capital stock's evolution at every moment, capturing the effects of investment and depreciation instantaneously. This provides a more precise and realistic depiction of gradual economic adjustments, especially when changes are ongoing rather than occurring in abrupt steps.

The rationale for adopting continuous-time frameworks arises from several key consid-

erations. For some research problems—particularly those involving high-frequency dynamics, instantaneous decision-making, or complex intertemporal optimization—continuous-time models can substantially simplify the mathematics. Concepts such as instantaneous rates of change, derivatives, and integral calculus become central, often leading to elegant closed-form solutions that might be difficult or impossible to obtain in a discrete-time setting. This mathematical elegance, in turn, can foster deeper intuition and facilitate a more precise characterization of economic mechanisms.

9.1 Basic tools and notation in continuous time

To begin, let's clarify how we represent variables in continuous time. In discrete-time models, we use subscripts to indicate time—for example, X_t for $t = 0, 1, 2, \ldots$. In continuous-time models, however, we write variables as *functions* of time: $X(t)$, where t can take any non-negative real value ($t \in \mathbb{R}_+$). This change reflects the fact that, in continuous time, variables can evolve at every instant, not just at fixed intervals.

The key mathematical concept in continuous-time analysis is the *time derivative*, which measures the instantaneous rate of change of a variable. For a function $X(t)$, the time derivative is written as $dX(t)/dt$ or, more compactly, $\dot{X}(t)$. This "dot notation," introduced by Newton (and hence also called "Newton notation"), is commonly used in economics to denote time derivatives.

Let's start by connecting these ideas to something familiar: growth rates. In discrete time, if X_t grows at a constant rate γ, we can write the relationship as a *difference equation*:

$$\frac{X_{t+1} - X_t}{X_t} = \gamma, \tag{9.1}$$

which has the solution

$$X_t = (1 + \gamma)^t X_0. \tag{9.2}$$

A difference equation relates the change in a variable from one period to the next to its current level, and its solution is a sequence of values at discrete points in time. In continuous time, we express the same idea using a *differential equation*. If the *instantaneous* growth rate of $X(t)$ is γ, we write:

$$\frac{\dot{X}(t)}{X(t)} = \gamma. \tag{9.3}$$

Here, $\dot{X}(t)$ is the instantaneous rate of change of $X(t)$, and dividing by $X(t)$ gives the proportional (or percentage) rate of change at each moment. The solution to this differential equation can be written using the exponential function:

$$X(t) = e^{\gamma t} X(0), \tag{9.4}$$

where e is Euler's number (approximately 2.71828). In this form, $X(t)$ grows smoothly at rate γ for every instant $t \geq 0$. You can check this solution by substituting it back into equation (9.3). This can be verified by differentiating (9.4) with respect to time: using the properties of exponential functions and their derivatives (see Appendices 9.A.1 and 9.A.2),

we have $\dot{X}(t) = \gamma e^{\gamma t} X(0) = \gamma X(t)$, which confirms that (9.4) solves the differential equation (9.3). It is important to recognize that the difference between discrete and continuous time extends to both notation and the types of equations used to describe dynamics. Difference equations (discrete time) generate sequences, while differential equations (continuous time) generate continuous functions of time.

A particularly useful property in continuous time is that the growth rate of $X(t)$ can be found by taking the natural logarithm and then differentiating (see Appendices 9.A.1 and 9.A.2):

$$\frac{\dot{X}(t)}{X(t)} = \frac{d}{dt} \log(X(t)). \tag{9.5}$$

This tells us that the derivative of the log of a variable gives its (instantaneous) growth rate. Intuitively, if a variable grows at a constant rate, its logarithm traces a straight line over time. This is why economists often plot time series—such as GDP or capital—on a log scale: constant growth appears as a straight line, making trends easier to spot. As an example, see Figure 2.1.

This property is is particularly powerful when dealing with functions that involve products or powers, which are common in macroeconomics. For example, consider the Cobb-Douglas production function:

$$Y(t) = z(t)K(t)^{\alpha} L(t)^{1-\alpha},$$

where $Y(t)$ is output, $z(t)$ is total factor productivity (TFP), $K(t)$ is capital, and $L(t)$ is labor. Taking logs,

$$\log(Y(t)) = \log\left(z(t)K(t)^{\alpha} L(t)^{1-\alpha}\right) = \log(z(t)) + \alpha \log(K(t)) + (1-\alpha)\log(L(t)),$$

where the second equality uses properties of natural logs (see Appendix 9.A.1). Taking the time derivative on both sides and applying (9.5) gives:

$$\frac{\dot{Y}(t)}{Y(t)} = \frac{\dot{z}(t)}{z(t)} + \alpha \frac{\dot{K}(t)}{K(t)} + (1-\alpha)\frac{\dot{L}(t)}{L(t)}.$$

This is the continuous-time version of the growth accounting equation (2.1) in Chapter 2. Here, each term represents the instantaneous growth rate of a component of output. This formula is exact under the Cobb-Douglas assumption.

The formula (9.5) can also be used to derive (9.3) from (9.4). For example, if $X(t) = e^{\gamma t} X(0)$, taking logs gives $\log(X(t)) = \gamma t + \log(X(0))$. Differentiating with respect to t brings us back to γ, showing the connection between exponential growth and constant growth rates. Similarly, in discrete time, taking logs and differences of $X_t = (1+\gamma)^t X_0$ reveals that the growth rate is approximately γ when γ is small.[1]

[1] It's also helpful to see how these continuous-time formulas relate to the discrete-time case. For example, if $X_t = (1+\gamma)^t X_0$. Taking logs of both sides delivers $\log(X_t) = t\log(1+\gamma) + \log(X_0)$. Computing the difference, we obtain $\log(X_{t+1}) - \log(X_t) = \log(1+\gamma)$. Because this equation can be rewritten as $\log\left(1 + \frac{X_{t+1}-X_t}{X_t}\right) = \log(1+\gamma)$ and $\log(1+x) \approx x$ for a small x, it is approximately the case that $\frac{X_{t+1}-X_t}{X_t} = \gamma$.

9.2 Optimization in continuous-time models: Maximum principle

Having established the key notation and mathematical tools, we now turn to optimization in continuous time. In this section, we introduce the maximum principle, also known as the optimal control or Hamiltonian method. To motivate the continuous-time approach, we begin with a familiar benchmark: the finite-horizon consumption-saving problem, similar to Section 4.1.1.

The consumer solves the problem:

$$\max_{\{c_t, a_{t+1}\}_{t=0}^{T}} \sum_{t=0}^{T} \beta^t u(c_t),$$

subject to

$$a_{t+1} = w + (1+r)a_t - c_t, \qquad a_0 \text{ given}, \tag{9.6}$$

and a terminal (no-borrowing) condition

$$a_{T+1} \geq 0. \tag{9.7}$$

Here, $\beta \in (0,1)$ is the discount factor, $u(\cdot)$ is the utility function that is strictly increasing, strictly concave utility function satisfying Inada conditions, c_t is consumption at period t, a_t is an asset at the beginning of period t, and $T > 0$ is the final period. The labor earnings w and interest rate r are given to the consumer.

We form a Lagrangian, slightly different from Section 4.1.1, as

$$L \equiv \sum_{t=0}^{T} \beta^t u(c_t) + \sum_{t=0}^{T} \mu_t((1+r)a_t + w - c_t - a_{t+1}) + \lambda a_{T+1},$$

where μ_t $(t = 0, 1, ..., T)$ are Lagrange multipliers (or costate variables) and λ enforces the terminal constraint. As in Section 4.1.1, we can proceed by applying the Kuhn-Tucker theorem to this problem directly, but here, we take a bit of a detour. The Lagrangian can be rewritten by expressing the budget constraint (9.6) in difference form,

$$a_{t+1} - a_t = ra_t + w - c_t. \tag{9.8}$$

Define the discrete-time *Hamiltonian* as

$$H_t \equiv \beta^t u(c_t) + \mu_t(ra_t + w - c_t).$$

Note H_t is a function of c_t and a_t (and μ_t). With this, the Lagrangian is:

$$L = \sum_{t=0}^{T} H_t - \sum_{t=0}^{T} \mu_t(a_{t+1} - a_t) + \lambda a_{T+1}.$$

Then, the first-order conditions for c_t, $t = 0, 1, 2, ..., T$ is

$$\frac{\partial H_t}{\partial c_t} = 0. \tag{9.9}$$

The first-order conditions for a_{t+1}, $t = 0, 1, 2, ..., T-1$ is

$$\frac{\partial H_{t+1}}{\partial a_{t+1}} + \mu_{t+1} - \mu_t = 0. \tag{9.10}$$

In this context, *control variables* (like c_t) are chosen at time t, while *state variables* (like a_t) capture the system's current position, carried over from previous choices. This distinction mirrors what was introduced in Chapter 4, Section 4.1. State variables cannot be changed at time t, but some state variables at period $t+1$ can be decided at period t. The control variable maximizes the Hamiltonian directly, while the state variable's condition accounts for changes in the costate variable μ_t over time (e.g., the fact that μ_t can change over time has to be taken into account). This consideration results in the extra term $\mu_{t+1} - \mu_t$.

The terminal condition (e.g., the FOC with respect to a_{T+1}) yields

$$-\mu_T + \lambda = 0. \tag{9.11}$$

The Kuhn-Tucker conditions are $\lambda \geq 0$, $a_{T+1} \geq 0$, and

$$\lambda a_{T+1} = 0. \tag{9.12}$$

Condition (9.9) for $t = T$ is $\beta^T u'(c_T) = \mu_T$, and combining this relationship with (9.11) and (9.12), we obtain

$$\beta^T u'(c_T) a_{T+1} = 0. \tag{9.13}$$

This condition is often called the *transversality condition* (TVC). We have discussed this concept in Chapter 4. In the finite-horizon case, because $\beta^T > 0$ and $u'(c_T) > 0$, the implication is that $a_{T+1} = 0$. This result does not necessarily apply when the planning horizon is infinite.

The condition (9.9) for time t is

$$\beta^t u'(c_t) = \mu_t,$$

and the condition (9.10) for time $t+1$ is

$$\mu_t = (1+r)\mu_{t+1}.$$

The standard Euler equation follows by combining FOCs:

$$u'(c_t) = \beta(1+r)u'(c_{t+1}). \tag{9.14}$$

Given the initial condition a_0, the budget constraint (9.6), and the TVC (9.13), this characterizes the optimal paths of c_t and a_t. Using the Euler equation repeatedly from time 0 to $T-1$,

$$\beta^T u'(c_T) = \frac{u'(c_0)}{(1+r)^T}$$

holds, and therefore (together with the fact that $u'(c_0)$ a positive number), (9.13) is equivalent to

$$\frac{a_{T+1}}{(1+r)^T} = 0.$$

Once again, in the finite-horizon case here, this condition is equivalent to $a_{T+1} = 0$.

For an infinite horizon, the setup is analogous:

$$\max_{\{c_t, a_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t),$$

subject to

$$a_{t+1} = w + (1+r)a_t - c_t,$$

a_0 is given, and

$$\lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} \geq 0. \tag{9.15}$$

We assume r is not too large, so that the present value of utility is finite. Note the last constraint looks somewhat different from the finite-horizon case (9.7). In the finite horizon, $a_{T+1}/(1+r)^T \geq 0$ is equivalent to $a_{T+1} \geq 0$ because $(1+r)^T > 0$ for any T, so it does not make a difference. In the infinite horizon, $\lim_{T\to\infty} a_{T+1}/(1+r)^T \geq 0$ is not equivalent to $\lim_{T\to\infty} a_{T+1} \geq 0$ (consider the case where $a_t = a > 0$ for all t; this satisfies the former inequality but not the latter inequality). See Section 4.3.1 for the discussion of why (9.15), called the no-Ponzi game (nPg) condition, is the appropriate one in this case.

As in the finite-horizon case, rewrite the budget constraint as $a_{t+1} - a_t = ra_t + w - c_t$ and define the Hamiltonian

$$H_t \equiv \beta^t u(c_t) + \mu_t(ra_t + w - c_t),$$

where μ_t is the Lagrange multiplier.

Then, the first-order conditions are

$$\frac{\partial H_t}{\partial c_t} = 0$$

for $t = 0, 1, ...$ and

$$\frac{\partial H_{t+1}}{\partial a_{t+1}} + \mu_{t+1} - \mu_t = 0.$$

These two equations can be combined to obtain the same Euler equation as in the finite-horizon case (9.14). The TVC is

$$\lim_{T\to\infty} \beta^T u'(c_T) a_{T+1} = 0.$$

This condition is analogous to the finite-horizon case (9.13). As in the finite-horizon case, this condition can also be expressed as

$$\lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} = 0.$$

See Section 4.3.1 for further discussions on this condition.

The continuous-time version is formulated by switching notation from X_t to $X(t)$ and replacing sums with integrals:

$$\max_{c(t), a(t)} \int_0^{\infty} e^{-\rho t} u(c(t)) dt$$

subject to

$$\dot{a}(t) = w + ra(t) - c(t), \tag{9.16}$$

$a(0)$ is given, and

$$\lim_{T\to\infty} e^{-rT} a(T) \geq 0. \tag{9.17}$$

Let's begin by examining the objective function. The period utility $u(\cdot)$ (often called the "instantaneous utility" or "moment utility" in the continuous-time setting) is identical to the one in discrete time. However, instead of summing utility across periods, we now take an integral to capture utility received at every instant in time. One key difference in the continuous-time formulation is how we discount future utility. Here, $\rho > 0$ is known as the *discount rate*, and discounting is accomplished using the exponential term $e^{-\rho t}$. This serves a similar purpose to β^t in discrete time: it reduces the weight of future utility relative to the present. Note that a larger discount rate ρ means the consumer places less value on future utility. By contrast, in discrete time, a larger discount factor β (closer to 1) means the consumer cares more about the future. For a formal explanation of the connection between ρ and β, see the box below.

The discount rate ρ

Suppose that, in a discrete-time setup, we discount the next period using the discount factor $\beta \in (0, 1)$. We define the *discount rate* ρ by the value that satisfies

$$\beta = \frac{1}{1+\rho}. \tag{9.18}$$

When we discount t periods ahead, we multiply β^t, or equivalently, $[1/(1+\rho)]^t$. Now, imagine we split one period into many subperiods, each of which has a length Δ. For example, when one period is one year, we can think of Δ being one month and $1/\Delta = 12$. Let us discount one subperiod by the rate $\rho\Delta$. Thus, when we discount a variable for t periods (which is equal to t/Δ subperiods), we multiply

$$f(\Delta, t) = \left(\frac{1}{1+\rho\Delta}\right)^{\frac{t}{\Delta}}.$$

When we take a limit $\Delta \to 0$,

$$\lim_{\Delta\to 0} f(\Delta, t) = \lim_{\Delta\to 0} \left(\frac{1}{1+\rho\Delta}\right)^{\frac{t}{\Delta}} = \lim_{\Delta\to 0} \left[(1+\rho\Delta)^{1/(\rho\Delta)}\right]^{-\rho t} = e^{-\rho t},$$

where we used the fact $e = \lim_{x\to 0}(1+x)^{1/x}$. Thus, the continuous-time discount rate ρ that shows up in $e^{-\rho t}$ can be interpreted as analogous to ρ in (9.18) when a period can be divided into many subperiods so that we can discount "continuously."

Constraint (9.16) is similar to (9.8), except that the change in the asset is expressed as $\dot{a}(t)$ instead of $a_{t+1} - a_t$. The nPg condition (9.17) is similar to (9.15), except that the

discounting is e^{-rT} instead of $1/(1+r)^T$. We can map these two with each other using the same procedure as in the above box. When r varies over time (denote it r_t or $r(t)$), we can generalize the nPg condition for the discrete-time case to

$$\lim_{T\to\infty} \frac{a_{T+1}}{\prod_{t=1}^{T}(1+r_t)} \geq 0$$

and the continuous-time case to

$$\lim_{T\to\infty} e^{-\int_0^T r(t)} a(T) \geq 0.$$

See Appendix 9.A.3 for detailed exposition.

Now, let us solve the problem. The steps are the same as in the discrete-time case. Define the Hamiltonian

$$H(t) \equiv e^{-\rho t} u(c(t)) + \mu(t)(ra(t) + w - c(t)), \tag{9.19}$$

where $\mu(t)$ is the costate variable. The first-order conditions become:

$$\frac{\partial H(t)}{\partial c(t)} = 0 \tag{9.20}$$

and

$$\frac{\partial H(t)}{\partial a(t)} + \dot{\mu}(t) = 0, \tag{9.21}$$

with the transversality condition

$$\lim_{T\to\infty} e^{-\rho T} u'(c(T)) a(T) = 0.$$

The only differences from the discrete-time case are that (i) the change in $\mu(t)$ is expressed as $\dot{\mu}(t)$ (instead of $\mu_{t+1} - \mu_t$) in (9.21), (ii) in (9.21), the relevant derivative of Hamiltonian is $\partial H(t)/\partial a(t)$ instead of $\partial H_{t+1}/\partial a_{t+1}$, because continuous-time setting has no "next period," and (iii) the discounting for the TVC is now $e^{-\rho T}$.

One can heuristically derive (9.20) and (9.21) in a similar manner as in the discrete-time case. Construct the Lagrangian

$$L = \int_0^\infty e^{-\rho t} u(c(t)) dt + \int_0^\infty \mu(t)(ra(t) + w - c(t) - \dot{a}(t)) dt.$$

(We ignore the terminal-condition issues here for the sake of exposition.) By rewriting the Hamiltonian as (9.19), the Lagrangian can be rewritten as

$$L = \int_0^\infty H(t) dt - \int_0^\infty \mu(t)\dot{a}(t) dt.$$

Applying integration by parts to the second term,

$$L = \int_0^\infty H(t) dt - \left[\lim_{T\to\infty} \mu(T)a(T) - \mu(0)a(0)\right] + \int_0^\infty \dot{\mu}(t)a(t) dt.$$

Taking the first-order conditions on $c(t)$ and $a(t)$ leads to (9.20) and (9.21).

Present-value Hamiltonian and current-value Hamiltonian

The Hamiltonian that is used in this chapter is often called the *present-value Hamiltonian*. An alternative formulation is called the *current-value Hamiltonian*. The current-value Hamiltonian evaluates the utility at the current value, without discounting. In the context of the above consumption-saving problem, the current-value Hamiltonian is

$$\hat{H}(t) \equiv u(c(t)) + \hat{\mu}(t)(ra(t) + w - c(t)).$$

With the current-value Hamiltonian, the first-order conditions are modified to

$$\frac{\partial \hat{H}(t)}{\partial c(t)} = 0 \tag{9.22}$$

and

$$\frac{\partial \hat{H}(t)}{\partial a(t)} + \dot{\hat{\mu}}(t) - \rho\hat{\mu}(t) = 0. \tag{9.23}$$

Equation (9.22) corresponds to (9.20), and (9.23) corresponds to (9.21). One can also see the relationship $\mu(t) = e^{-\rho t}\hat{\mu}(t)$.

Using (9.20) and (9.21), we can derive the continuous-time version of the Euler equation. First, (9.20) can be written as

$$e^{-\rho t}u'(c(t)) = \mu(t). \tag{9.24}$$

The first-order condition for the state variable, (9.21), can be calculated as

$$r\mu(t) + \dot{\mu}(t) = 0. \tag{9.25}$$

To eliminate $\mu(t)$, first rewrite (9.25) as

$$\frac{\dot{\mu}(t)}{\mu(t)} = -r.$$

Applying the growth trick (9.5) to (9.24),

$$-\rho + \frac{u''(c(t))}{u'(c(t))}\dot{c}(t) = \frac{\dot{\mu}(t)}{\mu(t)}.$$

Combining these two, we obtain

$$-\frac{u''(c(t))c(t)}{u'(c(t))}\frac{\dot{c}(t)}{c(t)} = r - \rho. \tag{9.26}$$

This equation is the continuous-time version of the Euler equation. The term $-u''(c(t))c(t)/u'(c(t))$ is called the *coefficient of relative risk aversion*, and because $u''(\cdot) < 0$, $c(t) > 0$, and $u'(\cdot) > 0$, the coefficient is always positive. In a class of utility functions, called the CRRA utility (introduced in Section 4.2.4), where $u(c) = (c^{1-\sigma} - 1)/(1 - \sigma)$, the coefficient is constant: $-u''(c(t))c(t)/u'(c(t)) = \sigma$.

Discrete-time Euler equation versus continuous-time Euler equation

Although the continuous-time Euler equation (9.26) and the discrete-time Euler equation (9.14) may look different, their economic interpretation and intuition are fundamentally the same. In the continuous-time case, equation (9.26) shows that the growth rate of consumption, $\dot{c}(t)/c(t)$, is determined by the balance between the benefit of saving (measured by the interest rate r) and the desire to consume sooner rather than later (measured by the discount rate ρ). If the consumer discounts the future heavily (i.e., ρ is large), then the growth rate of consumption will be low or even negative—meaning the consumer prefers to enjoy more consumption today. Conversely, when the discount rate ρ is small, the consumer is more willing to postpone consumption, leading to faster consumption growth over time.

A similar relationship is reflected in the discrete-time Euler equation (9.14). Here, a small discount factor β (corresponding to strong discounting) implies that marginal utility today, $u'(c_t)$, must be relatively low compared to marginal utility tomorrow, $u'(c_{t+1})$. This suggests that the consumer allocates more consumption to the present than to the future. To see the connection more clearly, consider a first-order Taylor expansion: $u'(c_{t+1}) = u'(c_t + [c_{t+1} - c_t]) \approx u'(c_t) + u''(c_t)(c_{t+1} - c_t)$ and recall that $\beta = 1/(1+\rho)$. Substituting this into (9.14), we obtain:

$$-\frac{u''(c_t)c_t}{u'(c_t)}\frac{c_{t+1}-c_t}{c_t} = \frac{r-\rho}{1+r},$$

which closely parallels the structure of the continuous-time Euler equation (9.26). This demonstrates that, despite the difference in notation, both approaches capture the same underlying trade-off between impatience and the rewards to saving.

9.3 Continuous-time growth models

In earlier chapters—specifically, Chapters 3 and 4—we introduced and analyzed the Solow model and the neoclassical growth model in discrete time. Those chapters provided the foundation for understanding long-run economic growth and dynamic optimization using the language of difference equations. Here, we revisit these classic models in continuous time. The purpose is twofold: first, to show how the same economic mechanisms can be formulated and analyzed using differential equations; and second, to provide hands-on examples that illustrate the tools and intuition developed earlier in this chapter. The continuous-time approach offers a more streamlined and, in some cases, more transparent characterization of the transition dynamics and steady states of growth models. For students, working through these continuous-time versions also clarifies the connections—and the distinctions—between discrete- and continuous-time techniques.

We begin with the continuous-time version of the Solow model, followed by the neoclassical growth model. In each case, we highlight both the economic interpretation and the formal differences that arise when moving from discrete to continuous time.

9.3.1 Solow model

Let us begin by revisiting the Solow model with economic growth, first introduced in Chapter 3.2. The Solow model remains a cornerstone of growth theory, illustrating how capital accumulation and technological progress drive the expansion of output per worker over time. By expressing the model in continuous time, we can see clearly how the economy's capital stock evolves at each instant, and how the interplay of savings, depreciation, population growth, and technological progress shape the long-run steady state.

The aggregate production function takes the form

$$Y(t) = F(K(t), A(t)L(t)),$$

where $Y(t)$ denotes aggregate output (GDP), $K(t)$ is aggregate capital, $A(t)$ represents the level of technology, and $L(t)$ is aggregate labor. The assumptions on the aggregate production function $F(\cdot, \cdot)$ are the same as in Chapter 3 (see Section 3.1). We assume that both labor and technology grow at constant rates:

$$\frac{\dot{L}(t)}{L(t)} = n \tag{9.27}$$

and

$$\frac{\dot{A}(t)}{A(t)} = \gamma. \tag{9.28}$$

The capital stock evolves according to the law of motion:

$$\dot{K}(t) = I(t) - \delta K(t), \tag{9.29}$$

where $I(t)$ is aggregate investment and $\delta > 0$ is the depreciation rate. As in Chapter 3, we assume that a constant fraction $s \in (0, 1)$ of aggregate income (which is the same as $Y(t)$) is saved and therefore invested:

$$I(t) = sF(K(t), A(t)L(t)). \tag{9.30}$$

To analyze the system's dynamics, it is helpful to express everything in terms of capital per effective unit of labor. Define

$$\tilde{k}(t) \equiv \frac{K(t)}{A(t)L(t)}.$$

Using the technique we learned in Section 10.1, we can obtain the growth rate of $k(t)$ as

$$\frac{\dot{\tilde{k}}(t)}{\tilde{k}(t)} = \frac{\dot{K}(t)}{K(t)} - \frac{\dot{A}(t)}{A(t)} - \frac{\dot{L}(t)}{L(t)}. \tag{9.31}$$

From (9.27), (9.28), (9.29), and (9.30), we can rewrite (9.31) as

$$\frac{\dot{\tilde{k}}(t)}{\tilde{k}(t)} = \frac{sF(K(t), A(t)L(t)) - \delta K(t)}{K(t)} - \gamma - n.$$

The left-hand side $frac\dot{\tilde{k}}(t)\tilde{k}(t)$ represents the proportional rate of change in capital per effective worker. The first term on the right is the net rate at which new capital is being created relative to the current capital stock. The subtracted terms, γ and n represent the rates of technological progress and population growth, respectively. To simplify further, divide the numerator and denominator of the first term by $A(t)L(t)$. Then, using the constant-returns property of the production function $F(\cdot,\cdot)$, we have:

$$\frac{F(K(t),A(t)L(t))}{A(t)L(t)} = F\left(\frac{K(t)}{A(t)L(t)},1\right) = f(\tilde{k}(t)),$$

where the first equality uses the constant-returns assumption on $F(\cdot,\cdot)$, and the second equality uses the definition $f(k) \equiv F(k,1)$ from Chapter 3. Substituting these definitions, we obtain

$$\frac{\dot{\tilde{k}}(t)}{\tilde{k}(t)} = \frac{sf(\tilde{k}(t))}{\tilde{k}(t)} - (\delta+\gamma+n).$$

Multiplying both sides by $\tilde{k}(t)$, yields a differential equation for the evolution of capital per effective worker[2]:

$$\dot{\tilde{k}}(t) = sf(\tilde{k}(t)) - (\delta+\gamma+n)\tilde{k}(t). \tag{9.32}$$

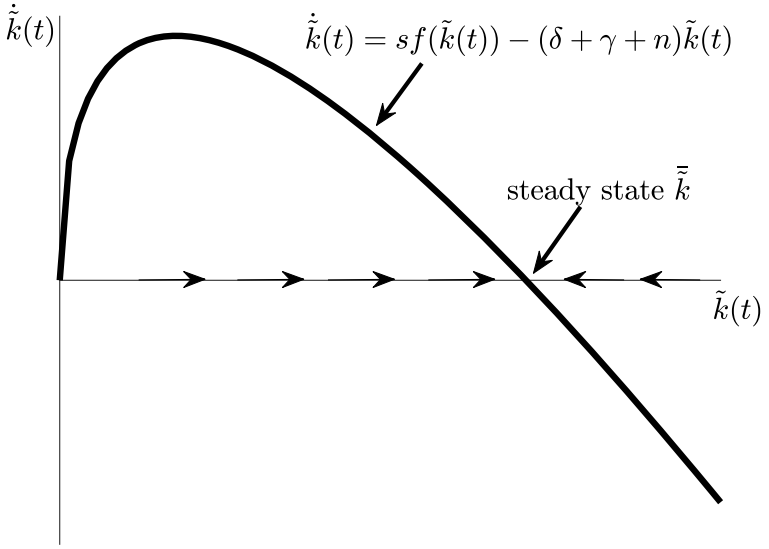


Figure 9.1: Solow model in continuous time

The dynamics described by equation (9.32) can easily be analyzed using the diagram depicted in Figure 9.1 (which is the analogous to Figure 3.1 in the discrete-time version). The vertical axis measures the rate of change of capital per effective worker, $\dot{\tilde{k}}(t)$—that is, the change of $\tilde{k}(t)$ over time— , while the horizontal axis shows its current level of capital

[2]This expression corresponds to the difference equations (3.5) and (3.7) in Chapter 3.

per effective worker, $\tilde{k}(t)$. The curve plotted in the figure represents the right-hand side of equation (9.32), incorporating the variables that drive the evolution of capital over time. The point where the curve crosses the horizontal axis corresponds to the steady state, denoted by $\bar{\tilde{k}}$. At this point, $\dot{\tilde{k}}(t) = 0$, implying that capital per effective worker remains unchanged over time. The behavior of the economy around this steady state reveals an important property of the Solow model: *convergence*. When $\tilde{k}(t)$ is to the left of the steady state ($\tilde{k}(t) < \bar{\tilde{k}}$), the value of $\dot{\tilde{k}}(t)$ is positive. This means that capital per effective worker is increasing, so the economy moves to the right in the diagram, gradually approaching the steady state. Conversely, when $\tilde{k}(t)$ exceeds the steady-state value ($\tilde{k}(t) > \bar{\tilde{k}}$), the rate of change $\dot{\tilde{k}}(t)$ becomes negative. In such case, caiptal per effective worker declines, and the economy moves to the left, once again approaching the steady state.

This dynamic adjustment process means that regardless of the economy's initial position, it will tend to move toward the steady-state level of capital per effective worker over time: from either direction, $\tilde{k}(t)$ gradually approaches $\bar{\tilde{k}}$. In the long run, the economy approaches the steady state, regardless of where it starts. The steady state itself is sometimes referred to as the balanced growth path, because along this path, the key economic variables—such as output per worker and capital per worker—grow at constant rates, determined by the rate of technological progress. In the long run, as the economy converges to the steady state, output per worker increases at the same rate as technology, and the overall path of the economy becomes predictable and stable. For example, because the per-capita GDP, $Y(t)/L(t)$, can be rewritten as

$$\frac{F(K(t), A(t)L(t))}{L(t)} = f(\tilde{k}(t))A(t)$$

and $f(\tilde{k}(t))$ is constant in the steady state, the growth rate of $Y(t)/L(t)$ in the long run is identical to the growth rate of $A(t)$, which is γ.

9.3.2 Neoclassical growth model

Let us now turn to the Neoclassical growth model, building on the analysis introduced in 4.1.2. To simplify the exposition, suppose there is no population growth or technological progress; these features can be incorporated using the same techniques as discussed previously and further developed in Section 4.3.3.

The planner's problem in continuous time

Consider the social planner's problem, formulated here in continuous time. The planner seeks to maximize the present discounted value of aggregate utility, choosing time paths for consumption and capital to solve

$$\max_{c(t),k(t)} \int_0^\infty e^{-\rho t} u(c(t)) dt,$$

subject to the dynamic resource constraint

$$\dot{k}(t) = f(k(t)) - \delta k(t) - c(t), \tag{9.33}$$

$$c(t), k(t) \geq 0, \quad \forall t,$$

and

$$k(0) \text{ given.}$$

Unlike in household problems with access to credit markets, the social planner in this setting cannot borrow against the future: the planner is bound by the non-negativity of capital at every instant (since this is a closed economy), so there is no need for an explicit no-Ponzi game condition. Following the same steps as in the previous consumption-saving problem, we can construct a Hamiltonian

$$H(t) \equiv e^{-\rho t} u(c(t)) + \mu(t)(f(k(t)) - \delta k(t) - c(t)).$$

where $\mu(t)$ is the costate variable associated with capital. The necessary first-order conditions are:

$$\frac{\partial H(t)}{\partial c(t)} = 0$$

$$\frac{\partial H(t)}{\partial k(t)} + \dot{\mu}(t) = 0.$$

In addition, the transversality condition ensures that resources are not wasted in the limit,

$$\lim_{T \to \infty} e^{-\rho T} u'(c(T)) k(T) = 0. \tag{9.34}$$

For further tractability, let us focus on the widely used CRRA utility function, $u(c) = (c^{1-\sigma} - 1)/(1-\sigma)$ with $\sigma > 0$ and $\sigma \neq 1$. Using the first-order conditions and following steps similar to those used in deriving equation (9.26) above, we can derive the Euler equation

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\sigma} \left(f'(k(t)) - (\delta + \rho) \right). \tag{9.35}$$

The evolution of the economy is governed by the system of two differential equations, (9.33) and (9.35), together with the initial condition $k(0)$ and the TVC (9.34). These equations jointly determine the dynamic paths of capital and consumption over time. The trajectory of $(k(t), c(t))$ can be analyzed with a phase diagram (similar to the one in the Appendix of Chapter 4), illustrated in Figure 9.2.

The horizontal axis corresponds to the stock of capital $k(t)$ and the vertical axis to consumption $c(t)$. From equation (9.33), the sign of $\dot{k}(t)$—which captures how the capital stock evolves—depends on the relationship between consumption and output net of depreciation. Specifically, capital gro ($\dot{k}(t) > 0$) whenever $c(t) < f(k(t)) - \delta k(t)$. In the phase diagram, the curve $c(t) = f(k(t)) - \delta k(t)$, shown as a dash-dot line, marks the boundary where capital is neither rising nor falling ($\dot{k}(t) = 0$). Below this line, where consumption is lower, capital increases over time (illustrated by rightward arrows). Conversely, above this line, higher consumption leads to a decline in capital. Turning to equation (9.35), we see that consumption rises over time ($\dot{c}(t) > 0$) only when $f'(k(t)) - (\delta + \rho) > 0$. This condition depends solely on the level of capital, $k(t)$. Because $f'(\cdot)$ function is decreasing, this inequality implies $\dot{c}(t) > 0$ if and only if $k(t) < k^{**}$, where k^{**} satisfies $f'(k^{**}) = \delta + \rho$. The dotted line, $\dot{c}(t) = 0$, represents $k(t) = k^{**}$. To the left of this line, $c(t)$ increases over time (represented by the

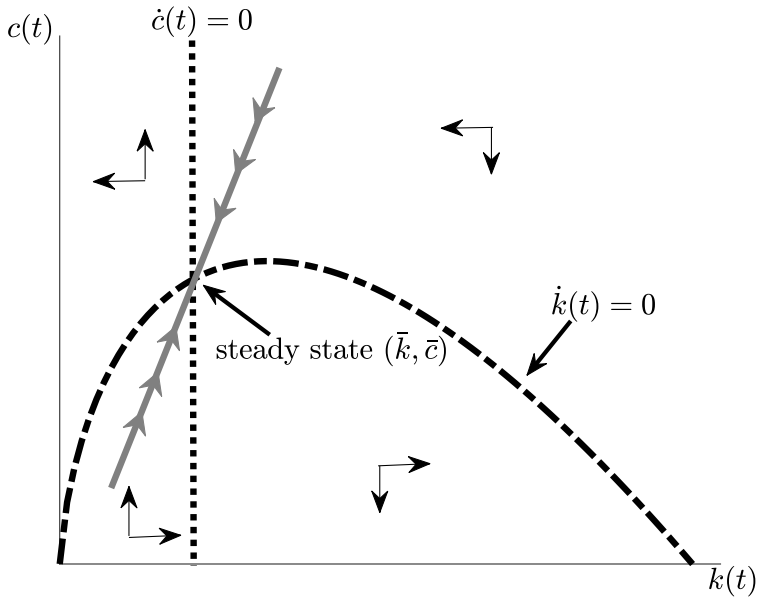


Figure 9.2: Phase diagram for Ramsey model in continuous time

upward arrows). Similarly, $c(t)$ decreases over time on the right side of the dotted line. The value k^{**} is called the *modified golden rule* capital stock. This value is smaller than the *golden rule* capital stock, which maximizes steady-state consumption.[3]

Starting from $k(0)$, the social planner has to choose the value of initial consumption $c(0)$ that is on the saddle path that is represented by the solid gray line. On the saddle path, $(k(t), c(t))$ gradually approaches the steady state $(\bar{k}, \bar{c})$ (here, $\bar{k} = k^{**}$) over time. This dynamic also satisfies the TVC (9.34), because $k(t)$ and $c(t)$ become constant in the long run. Thus, once again, the Ramsey model exhibits *convergence*. To see the other values of $c(0)$ cannot be chosen, consider first the case when a larger value of $c(0)$ is chosen. One can see from the phase diagram that eventually $(k(t), c(t))$ reaches the situation $k(t) = 0$. At that point, the economy cannot produce anything, which implies $c(t)$ following the Euler equation is not feasible. Thus, this value of $c(0)$ does not satisfy the optimality conditions. If $c(0)$ is smaller than the one on the saddle path, eventually $(k(0), c(0))$ reaches the region where the value of $k(t)$ is large and $c(t) \to 0$. Appendix 9.A.4 shows this path does not satisfy the TVC. The consumers over-accumulate capital and do not consume much, which is a waste, as can be seen from the fact that the path of $c(t)$ is lower than the one on the saddle path.

[3]In Figure 9.2, the golden rule capital stock is the value of k that maximizes the dash-dot curve, since this curve represents steady-state consumption $c = f(k) - \delta k$. The maximizing condition, $f'(k) = \delta$, defines the golden rule. In contrast, the modified golden rule capital stock,k^{**}, solves $f'(k^{**}) = \delta + \rho$. Because $\rho > 0$, k^{**} is smaller than the golden rule level: consumers' impatience leads to less capital accumulation in the steady state.

The market equilibrium in continuous time

The analysis above characterizes the Pareto-optimal allocation, where all individuals are treated identically by the social planner. We now show that this allocation coincides with the outcome of the market equilibrium. While this equivalence follows from the first and second welfare theorems, it is nevertheless instructive to see explicitly how the decentralized market solution mirrors the planner's problem in this setting—offering another application of continuous-time methods.

Consider a continuum of households indexed by $i \in [0,1]$, each with identical preferences and the same initial capital stock $k_i(0)$. The problem faced by a representative household is:

$$\max_{c_i(t),k_i(t)} \int_0^\infty e^{-\rho t} u(c_i(t))dt,$$

subject to

$$\dot{k}_i(t) = r(t)k_i(t) + w(t) - \delta k_i(t) - c_i(t), \tag{9.36}$$

$$c_i(t), k_i(t) \geq 0, \quad \forall t,$$

and

$$k_i(0) \text{ given.}$$

Here, $r(t)$ denotes the rental rate of capital and $w(t)$ the wage rate. Applying the same optimization steps as before, the household's problem delivers the Euler equation

$$\frac{\dot{c}_i(t)}{c_i(t)} = \frac{1}{\sigma}\left(r(t) - (\delta + \rho)\right). \tag{9.37}$$

The representative firm maximizes profit. The firm's problem is static:

$$\max_{K(t),L(t)} \; F(K(t), L(t)) - r(t)K(t) - w(t)L(t),$$

where $K(t)$ is the factor demand for capital and $L(t)$ is the labor demand. The first-order conditions are

$$F_1(K(t), L(t)) = r(t)$$

and

$$F_2(K(t), L(t)) = w(t),$$

where $F_i(K(t), L(t))$ is the partial derivative with respect to term i. As in Section 3.3, we can show $F_1(K(t), L(t)) = f'(\tilde{k}(t))$ and $F_2(K(t), L(t)) = f(\tilde{k}(t)) - \tilde{k}(t)f'(\tilde{k}(t))$, where $f(k) \equiv F(k, 1)$ and $\tilde{k}(t) \equiv K(t)/L(t)$.

Let us now turn to the market equilibrium. There are three markets in this economy: the product market, the capital (rental) market, and the labor market. By Walras' law, we only need to consider two markets. In the capital market, the market-clearing condition is

$$\int k_i(t)di = K(t).$$

The right-hand side is the labor demand, and the left-hand side is the capital demand. In the labor market, because the aggregate supply of labor is 1, $1 = L(t)$. Let the aggregate capital in this economy be denoted as

$$k(t) = \int k_i(t)di.$$

Then, from the firm's optimization, the equilibrium prices are

$$r(t) = f'(k(t)) \tag{9.38}$$

and

$$w(t) = f(k(t)) - k(t)f'(k(t)). \tag{9.39}$$

Replacing (9.38) and (9.39) in (9.36) and (9.37) and using symmetry ($c_i(t) = c(t)$ and $k_i(t) = k(t)$), we obtain

$$\dot{k}(t) = f(k(t)) - \delta k(t) - c(t)$$

and

$$\frac{\dot{c}(t)}{c(t)} = \frac{1}{\sigma}\left(f'(k(t)) - (\delta + \rho)\right),$$

which are identical to (9.33) and (9.35).

9.4 Uncertainty in continuous-time settings: Poisson process

Macroeconomic environments are inherently uncertain. Throughout this text, we have seen how uncertainty shapes savings decisions, labor supply, investment, and policy—whether in the form of aggregate shocks (see Chapters 7 and 14), income risks (Chapter 11), or events such as innovation and sovereign default (Chapters 13 and 24). In discrete-time models, it is natural to introduce uncertainty as a "shock in each period." For example, by specifying that output, productivity, or income can jump to a new value each quarter or year, according to some probability distribution. However, continuous-time modeling, which is essential in many applications (such as those in Chapter 13), does not come with a natural unit of time. So, how should we represent uncertainty in a continuous world?

There are two broad approaches for incorporating randomness in continuous time. The first captures the notion of uncertainty as a constant flow of small, frequent shocks, leading to what is called a *diffusion process*. Brownian motion, the prototypical example, is widely used in finance and has become increasingly prominent in macroeconomics, particularly in the analysis of asset prices, precautionary saving, and risk. However, a rigorous treatment of diffusion processes requires mathematical tools beyond the scope of this book.

The second approach focuses on the idea that randomness manifests as rare but discrete, often substantial, changes—events that occur unpredictably at random points in time (e.g., both the size and the frequency of shocks can be random). This class of models is built on the mathematics of *jump process*, the most fundamental of which is the Poisson process. The Poisson process is especially useful for modeling situations where events—such

as technological breakthroughs, job loss, firm entry or exit,—occur sporadically but with a well-defined average frequency. This structure underlies many applications in macroeconomics. For instance, the Poisson process plays a central role in the analysis of job search and unemployment dynamics (see Chapter 20), and it forms the backbone of models of innovation and endogenous growth (see Chapter 13). By focusing on the Poisson process, we can introduce a rich form of uncertainty into continuous-time macroeconomic models in a way that remains mathematically accessible and tightly connected to a wide range of economic phenomena[4].

To consider a Poisson process, let us start with Bernoulli trials in discrete time. Consider a Bernoulli trial that can result in a "success" or a "failure." The situation can be finding a job, losing a job, succeeding in innovation, etc. Suppose the probability of success is $\lambda \in [0, 1]$ for one trial. Suppose that during each unit of time (say, a year), one trial is made. From these assumptions, the expected number of successes during this one year is λ. Now, suppose that we divide the period into two and have one trial every six months. Each trial is independent. If we adjust the success probability of each trial to $\lambda/2$, the expected total number of successes during the one year is still λ. In general, if we divide the period (one year) into n subperiods and make the success probability λ/n in each subperiod, we still keep the expected total number of successes λ, although now we may have many successes during one year. The expected number of successes can be computed as $\underbrace{\frac{\lambda}{n} + \cdots + \frac{\lambda}{n}}_{n \text{ times}} = \lambda$.

With n trials, the distribution of the number of successes follows a binomial distribution

$$b(k, n) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}, \tag{9.40}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is the binomial coefficient (the number of different ways one can have k successes out of n trials). $b(k, n)$ represents the probability of k successes during this time period.

Note that, in (9.40),

$$b(0, n) = \left(1 - \frac{\lambda}{n}\right)^n \tag{9.41}$$

and

$$\frac{b(k, n)}{b(k-1, n)} = \frac{n - (k-1)}{k} \frac{\lambda/n}{1 - \lambda/n} = \frac{\lambda - (k-1)\lambda/n}{k - \lambda k/n} \tag{9.42}$$

for $k \geq 1$ hold. Let us consider a situation where $n \to \infty$. That is, the length of each subperiod approaches zero. In (9.41),

$$p(0) \equiv \lim_{n \to \infty} b(0, n) = e^{-\lambda}. \tag{9.43}$$

The second inequality follows from the definition of e. Taking the limit of (9.42),

$$\lim_{n \to \infty} \frac{b(k, n)}{b(k-1, n)} = \frac{\lambda}{k}$$

[4]A portion of the exposition that follows draws on Feller (1968).

holds. Applying this formula for $k = 1, 2, 3, ...$ sequentially, we obtain

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

This $p(k)$ represents the probabilities for the *Poisson distribution*. That is, $p(k)$ is the probability of k successes when the number of trials n approaches infinity (and the probability of success in each trial, λ/n, approaches zero).

The *Poisson process* is a stochastic process where the distribution of the number of successes between any time interval $(t, t+T]$, for $T > 0$, follows a Poisson distribution with

$$p(k) = e^{-(\lambda T)}\frac{(\lambda T)^k}{k!}.$$

The expected number of successes during the time interval $(t, t+T]$ is λT, and we saw that the Poisson process in continuous time can be understood as the limit of the discrete-time situation where repeated independent Bernoulli trials are conducted in every small time interval $\Delta = T/n$, where $\Delta \to 0$. Each trial has the probability of success $\lambda\Delta$, and the total number of trials is $n = T/\Delta$, thus keeping the expected number of successes as λT during the entire time period. Because the probability of success in each trial $\lambda\Delta \to 0$ and the total number of trials $n = T/\Delta \to \infty$ as $\Delta \to 0$, successes occur "infrequently" relative to the total number of trials.

A few notes are in order. First, by construction, the outcomes during $(t, t+T]$ and $(t+T+s, t+T+s+S]$ are independent of each other for any $s \geq 0$ and $S \geq 0$. Thus, the Poisson process is a memoryless process. Second, the initial restriction $\lambda \leq 1$ is not necessary; we can always start from a small-enough time interval (a large-enough n) such that $\lambda/n \leq 1$, and proceed with the above construction. Third, the probability that no success occurs during the time interval $(t, t+T]$ is, from (9.43), $e^{-\lambda T}$. If λ is time variant (denote as $\lambda(s)$ for time s), the same probability is expressed by $e^{-\int_t^{t+T}\lambda(s)ds}$.

As an example, consider the dynamics of employment and unemployment for a group of workers. Suppose that, for an individual employed worker, the event of job separation follows a Poisson process with parameter $\sigma > 0$. Then, the probability of keeping the job up to time t (i.e., the probability that no separation occurs up to time t) is, from the above formula, $e^{-\sigma t}$. Because the probability of separation during the time interval of length dt starting from time t is $e^{-\sigma t} \times \sigma dt$ (the probability of keeping the job up to time t times the probability of a separation event during the time period dt), the expected length of the remaining duration at the job is

$$\int_0^\infty te^{-\sigma t}\sigma dt = \frac{1}{\sigma}.$$

Consider a group of many workers with total population N. Let the number of employed worker at time t be $e(t)$ and the number of unemployed worker be $u(t)$. Assume that a worker is either employed or unemployed, that is, $e(t) + u(t) = N$. Also assume that during any (short) time interval dt, an unemployed worker finds a job with probability λdt, where $\lambda > 0$. The total number of new job matches during dt time interval is $u(t) \times \lambda dt$ from the

law of large numbers. Similarly, $e(t) \times \sigma dt$ is the total number of separation. Thus $u(t)$ follows the differential equation

$$du(t) = e(t)\sigma dt - u(t)\lambda dt,$$

where $du(t)$ is the change of $u(t)$ during the dt time interval, or, using the dot notation and $e(t) = N - u(t)$,

$$\dot{u}(t) = (N - u(t))\sigma - u(t)\lambda. \tag{9.44}$$

The steady-state mass of unemployment, $\bar{u}$, is (by setting $\dot{u}(t) = 0$)

$$\bar{u} = \frac{\sigma N}{\sigma + \lambda}.$$

By examining (9.44) closely, one can see the unemployment converges to the steady-state value, regardless of the initial unemployment. The unemployment rate in the steady state is

$$\frac{\bar{u}}{N} = \frac{\sigma}{\sigma + \lambda},$$

which is increasing in σ and decreasing in λ. A similar dynamics with discrete time (and $N = 1$) is presented in Chapter 20.

Chapter 10

Computational tools

Lucas (1980) famously wrote "Our task as I see it...is to write a FORTRAN program that will accept specific economic policy rules as 'input' and will generate as 'output' statistics describing the operating characteristics of time series we care about, which are predicted to result from these policies." We do this work using computational tools rather than paper and pencil because, in many cases, macroeconomic models do not have tractable analytical solutions. This chapter introduces some core tools that underpin these computational approaches. The goal here is to give the reader a bird's eye view of these methods and how they fit together within the macroeconomist toolkit.[1]

The chapter starts with some building block methods, which are then combined to solve a dynamic programming problem in Sections 10.5. In practice, one can easily find software that performs the building block methods described below and one would not program these methods oneself. But understanding how these methods work is important in choosing which method to apply or diagnosing a problem when an algorithm does not behave as expected. Section 10.6 presents a perturbation method and is self-contained.

10.1 Approximating a function

We often encounter a situation where we need to store the information of a function in the computer. It is easy in some cases. For example, the information of a polynomial $f(x) = ax^2 + bx + c$ can be summarized by three numbers: a, b, and c. In many situations in macroeconomics, however, the function of interest does not have such a simple form. For example, when we solve Bellman equations in dynamic programming, most of the time, the value functions and policy functions do not have easy analytical expressions. Storing the value of the function at each point on an interval $[\underline{x}, \overline{x}]$ would in principle require infinite numbers, which is infeasible with finite computer memory. So, we need some way to approximate a function and this section introduces two popular methods.

[1] Readers looking for more details of these methods should see Numerical Recipes `http://numerical.recipes/book.html`, QuantEcon `https://quantecon.org/`, Judd (1998), Miranda and Fackler (2002).

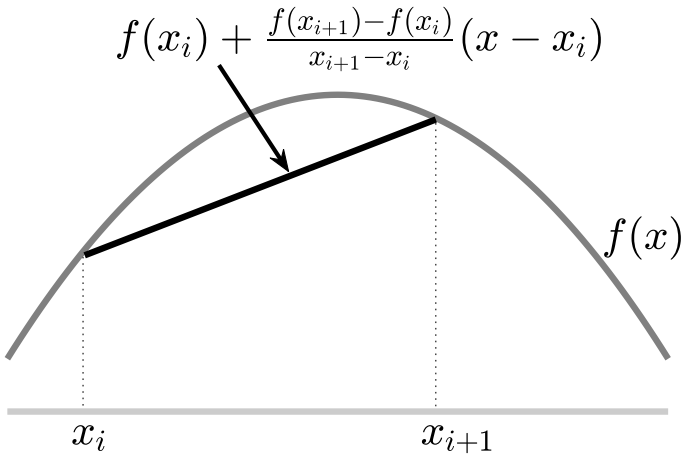


Figure 10.1: Linear interpolation

10.1.1 Interpolation

One natural method of approximation is to set discrete grid points on $[\underline{x}, \overline{x}]$ and store the functions value on these points. Let us index the grid points by $i \in \{1, ..., n\}$. It is natural to set $x_1 = \underline{x}$ and $x_n = \overline{x}$. In between, it is natural to have equally spaced points, but one might instead put more points in a subinterval where a more accurate approximation to the function is desired.

Linear interpolation

The most straightforward interpolation method is the linear interpolation. It only requires storing the values of $f(x_i)$ for all i. Then, given any value of x, we can find an approximate value of $f(x)$ by

1. Find the i where $x \in [x_i, x_{i+1})$.
2. Compute the approximated value $\hat{f}(x)$ by

$$\hat{f}(x) = f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}(x - x_i).$$

   As we can see from Figure 10.1, this formula provides the line that connects $(x_i, f(x_i))$ and $(x_{i+1}, f(x_{i+1}))$.

This method has the advantage of using only local information ($f(x_{i+1})$ and $f(x_i)$) and is thus robust to what is happening elsewhere. The two disadvantages are that (i) the approximation error can be large if the underlying function is highly nonlinear, and (ii) the approximated function is not differentiable at the grid points.

Cubic spline interpolation

Another popular method is cubic spline interpolation, which can often achieve a better approximation of the underlying function than linear interpolation. Moreover, the approxi-

mated function is twice continuously differentiable. The latter property is often important in economics.

The idea is to interpolate using a (particular kind of) cubic function instead of a linear function. Then, we require the cubic function to have the same first and second derivatives from the right and from the left at each of the grid points $x_2, ..., x_{n-1}$. We are able to devise a cubic function that only requires $f(x_i)$, $f(x_{i+1})$, $f''(x_i)$, and $f''(x_{i+1})$ to interpolate between x_i and x_{i+1}. Moreover, we do not need the actual information of the second derivatives—we can compute them from the requirement that the first derivatives are continuous at each grid point (plus some boundary conditions). Thus, once again, the only information we have to store is the values of $f(x_i)$ for all i. The downside compared with the linear interpolation is that step 1 uses the information of all $f(x_i)$s to compute a particular $f''(x_i)$, and thus, in principle, the value of $f(x_j)$ could affect the interpolation in $x \in [x_i, x_{i+1})$ even when x_j is far away from x_i.

10.1.2 Approximation by known functions

The second popular method is to use functions that are known to the computer (e.g., such as polynomials, log functions, and trigonometric functions) and consider their "weighted sum":

$$\hat{f}(x) = \sum_{i=1}^{n} a_i \phi_i(x). \tag{10.1}$$

Here, we choose n different functions $\phi_i(x)$, and the weights are a_i. A typical method of setting a_i is to choose various points in x, $\{x_1, ..., x_m\}$ and use the information of $\{f(x_1), ..., f(x_m)\}$ in addition to $\{\phi_i(x_1), ..., \phi_i(x_m)\}$ for all i. If $m = n$, (generically) a unique set of a_is solve (10.1) when the left-hand side is set as $f(x_j)$ (m equations with n unknowns). If $m \geq n$, we cannot set a_is to set (10.1) to hold with equality, but we can use (10.1) as a regression equation to choose the regression coefficients a_is so that the right-hand side is "close" to the left-hand side.

A popular choice of ϕ_is is a set of functions called the Chebychev polynomials.[2] The Chebychev polynomials are known to summarize the information of $f(x)$ in an efficient manner, and the methods of choosing $\{x_1, ..., x_m\}$ to find the a_i's are also well established. The downside of this method is similar to the cubic spline interpolation. Because this approximation method uses global information, an approximation of one part is influenced by the behavior of $f(x)$ and $\phi_i(x)$ in other parts of the domain.

10.2 Root finding

In economics, we often have to solve for the root of a nonlinear equation. Here, we focus on the one-dimensional case, that is, finding a scalar x that satisfies

$$f(x) = 0.$$

[2]Codes for implementing Chebychev polynomial approximation can be found, for example, in the Numerical Recipes.

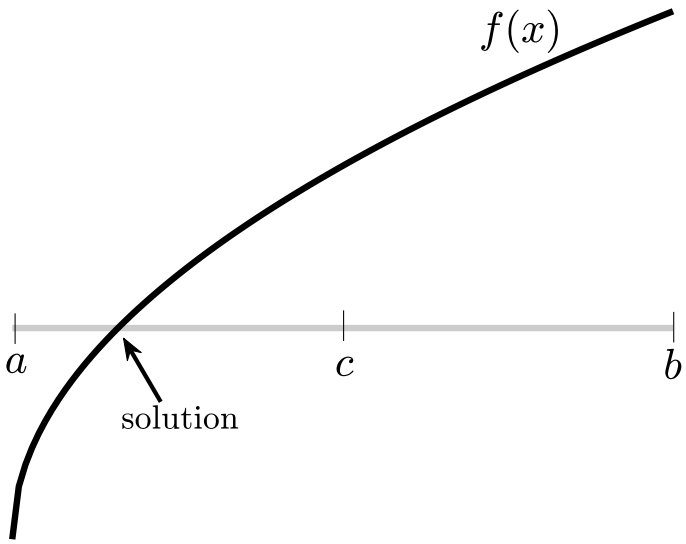


Figure 10.2: Bisection method

Methods for root finding in multi-dimensional cases often rely in similar principles.

The most obvious method of finding a root brute-force grid search. If we know a root of $f(x) = 0$ exists within an interval $[\underline{x}, \overline{x}]$, create grid points $\{x_1, ..., x_n\}$ between $x_1 = \underline{x}$ and $x_n = \overline{x}$ for a large n. Evaluate $f(x_i)$ for each i and pick x_i that corresponds to $f(x_i)$ closest to zero as the solution. This method always works and provides a good solution when n is a very large number. However, this method is extremely slow and is rarely the best choice.

10.2.1 Bisection

Bisection is related to brute-force grid search, but it is a lot more efficient. The steps are the following:

1. Find a and $b > a$ where $\text{sign}(f(a)) \neq \text{sign}(f(b))$. For example, in Figure 10.2, $f(a) < 0$ and $f(b) > 0$. This step is called "bracketing." If $f(x)$ is continuous, this condition implies that at least one solution to $f(x) = 0$ exists in the bracket $[a, b]$.

2. Let $c \equiv (a+b)/2$. Evaluate $f(c)$. If $\text{sign}(f(c)) = \text{sign}(f(a))$, a solution is in the bracket $[c, b]$ and we update the value of a to $a = c$. Otherwise, we update the value of b as $b = c$. In Figure 10.2, $\text{sign}(f(c)) = \text{sign}(f(b))$, and thus the solution is in $[a, c]$, and thus c becomes the new b.

3. Repeat step 2 until $(b - a)$ is less than a tolerance value you set (e.g., 10^{-6}). Check that $f(a)$ is close to zero, and use a as the solution.

The bisection method has several benefits. For one, it does not require differentiability of $f(x)$. In addition, even when $f(x)$ takes zero multiple times, this procedure always finds one of the solutions provided that the function is continuous. A big advantage of bisection is that it always ends after a set time. As each iteration cuts the interval in half, we know how many iterations it requires to shrink the starting interval down the tolerance level. The procedure

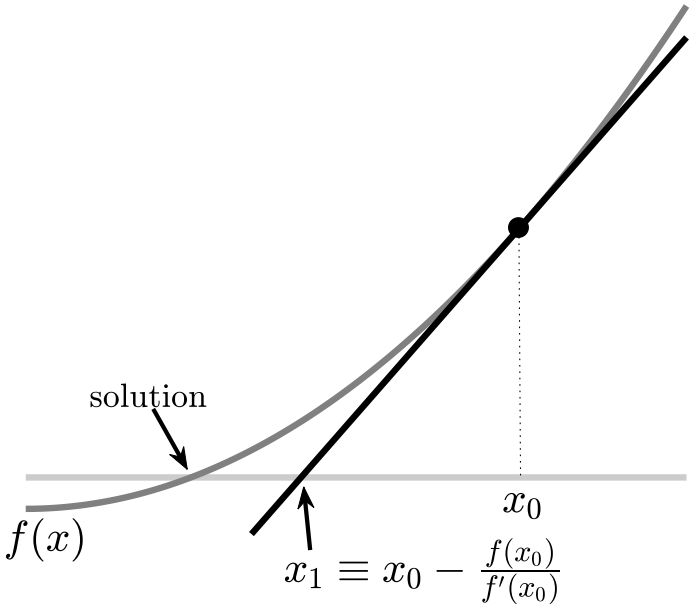


Figure 10.3: Constructing x_1 in the Newton-Raphson method

always ends, even when no solutions exist, for example, because $f(x)$ is not continuous. If we don't know the properties of $f(x)$, checking $f(a) \approx 0$ in the final step is therefore very important.

The bisection procedure is much faster than the brute-force grid search method described earlier. Suppose we are trying to find a solution of $f(x) = 0$ on $[0, 1]$ with 10^{-3} accuracy in terms of x. This process requires 1000 evaluations of $f(x)$ in the case of the brute-force grid search, but only 10 evaluations in the case of bisection (because $(1/2)^{10} = 1/1024$). This method can be slower than other methods (e.g., the following Newton-Raphson method can be much faster if $f(x)$ is close to linear), but it is quite a robust and useful method despite its simplicity.

10.2.2 Newton-Raphson

The idea of the Newton-Raphson method is based on linear approximation of a function:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

for some starting point x_0. If this approximation is good, the solution for $f(x) = 0$ should, by plugging 0 into the left-hand side, be close to

$$\hat{x} = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{10.2}$$

If $f(x)$ is indeed linear, $\hat{x}$ delivers the solution in one step. The general procedure is as follows:

1. Make an initial guess x_0.

2. Construct x_1 from

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Figure 10.3 describes how x_1 is constructed.

3. Check whether $f(x_1) \approx 0$ using some tolerance value. That is, if $|f(x_1)| < \varepsilon$, where ε is the tolerance value, stop and call x_1 the solution. If not, now use x_1 and obtain x_2, going back to the previous step. Continue until $f(x_i) \approx 0$.

If $f(x)$ is close to linear, or x_0 is close to the true solution, this procedure can be much faster than bisection. The downside is that (i) it requires differentiability of $f(x)$, and we need to compute $f'(x)$ either analytically or numerically, and (ii) even when the solution exists, the procedure may not be able to find a solution if the function is not well behaved.

One method of finding a numerical derivative is to compute a finite-difference derivative

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}.$$

In principle, a smaller h would provide a better approximation. In practice, because a computer does not recognize a value smaller-than-certain value (often called the "machine epsilon," typically around 10^{-16}), setting h too small is not desirable. A typical value used for h is around 10^{-5}.

10.3 Optimization

Let us consider maximizing a function $F(x)$. Similarly to the root finding, we can, in principle, optimize $F(x)$ by a simple grid search: evaluate $F(x)$ on many grid points and pick the point x that yields the maximum value of $F(x)$. This method is, although fail-safe, very slow. Below, we introduce two popular methods.

We focus on the one-dimensional case where x is a scalar. For multi-dimensional cases, one approach is to optimize one dimension at a time. When maximizing $F(x, y)$, for example, fix x and maximize $F(x, y)$ in terms of y (this is a one-dimensional problem). By performing this procedure for many x, we can obtain the function of maximizing y: $y^*(x)$. Then, we can perform a one-dimensional problem of maximizing $F(x, y^*(x))$ over x.

10.3.1 Golden-section search

The first method, as in the bisection method in the previous section, only requires evaluating the values of the function. The procedure is as follows.

1. First, find three values a, b, and c such that $a < b < c$ and $F(a) < F(b)$ and $F(c) < F(b)$. These properties mean that a (local) maximum exists between a and c. (If such values cannot be found, a corner solution is likely: either a or c is the maximum.) This set of the three points (a, b, c) is the initial "bracket."

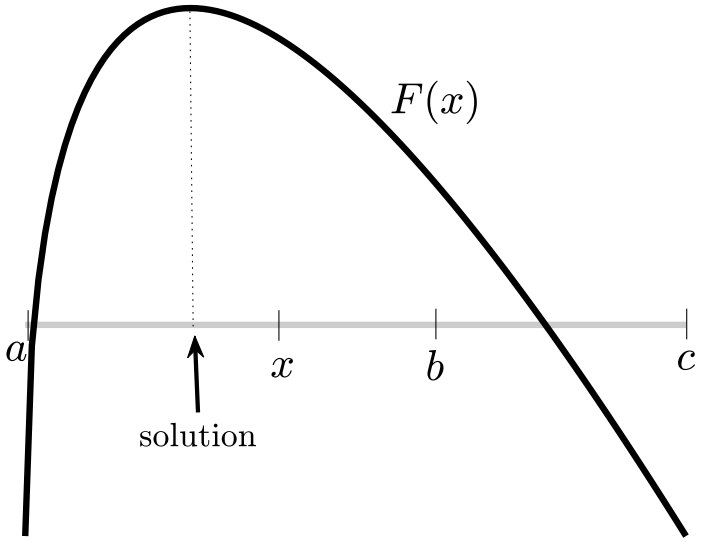


Figure 10.4: Golden section search

2. Take the longer segment between $[a, b]$ and $[b, c]$. Here, for exposition, suppose $(b - a) > (c - b)$; that is, $[a, b]$ is the longer one. Then, take a point x in $[a, b]$ so that $(b - x)/(b - a) = \omega$. Here, $\omega = (3 - \sqrt{5})/2 \approx 0.38197$. As a result, we have now broken our initial interval up into three segments $[a, x]$, $[x, b]$, and $[b, c]$. Now compare $F(x)$ and $F(b)$. If $F(x) > F(b)$, remove the segment $[b, c]$ and create the new bracket (a, x, b). Otherwise, remove the segment $[a, x]$ and create the new bracket (x, b, c). In Figure 10.4, $F(x) > F(b)$, and thus the new bracket becomes (a, x, b).

3. Repeat step 2 until the size of the bracket is less than a tolerance value.

In step 2, we use the parameter ω. The ratio $(1-\omega)/\omega \approx 1.61803$ is often called the "golden ratio" or "golden section." This ratio has been studied mathematically since ancient Greece and is widely used in the context of architecture and art. Here, ω is used because if we start from $(c - b)/(c - a) = \omega$ and follow the above procedure, every time, we can remove the ω fraction of the segment. After iterating n times, therefore, the length of the remaining bracket is $(1 - \omega)^n(c - b)$.

As in the case of bisection, this method does not require differentiability of $F(x)$, and it ends in a pre-set number of iterations.

10.3.2 Newton's method

The second method starts by approximating the function with a quadratic equation:

$$F(x) \approx F(x_0) + F'(x_0)(x - x_0) + \frac{1}{2}F''(x_0)(x - x_0)^2.$$

Taking the first-order condition of the right-hand side with respect to x and equating it with zero, we obtain the solution for x:

$$\hat{x} = x_0 - \frac{F'(x_0)}{F''(x_0)}. \tag{10.3}$$

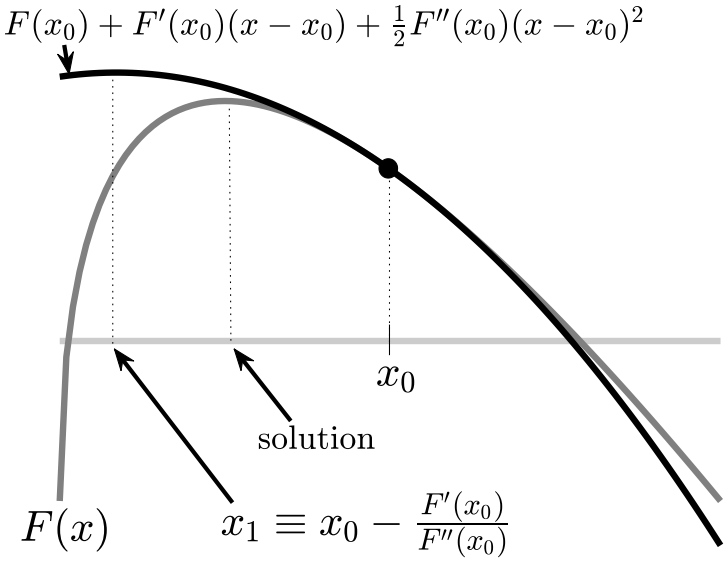


Figure 10.5: Constructing x_1 in the Newton method

If $F(x)$ is indeed quadratic, $\hat{x}$ delivers the solution in one step.

The general procedure is as follows:

1. Make an initial guess x_0.

2. Construct x_1 from
$$x_1 = x_0 - \frac{F'(x_0)}{F''(x_0)}.$$
See Figure 10.5 for an example.

3. Check whether $|x_1 - x_0|$ is less than the pre-set tolerance value ε. If it is indeed the case ($|x_1 - x_0| < \varepsilon$), stop and call x_1 the solution. If not, now use x_1 to obtain x_2, going back to the previous step. Repeat until $|x_{i+1} - x_i| < \varepsilon$.

Compared with the golden-section search, the Newton method can be substantially faster, especially if $F(x)$ is close to quadratic. However, the same issues as the Newton-Raphson method above can apply: it requires information on the first and second derivatives, and it is not guaranteed to find the maximum. Starting from a good first guess x_0 is helpful.

10.3.3 Connections between root finding and optimization

Newton's method is that it is the same as taking the first-order condition of a quadratic objective and finding its root by the Newton-Raphson method. The equivalence is clear from comparing equations (10.2) and (10.3), given that the first-order condition is $F'(x) = 0$.

In this section, we introduced two methods for one-dimensional optimization. As indicated above, in general, "optimizing" can be considered "finding a root of the first-order

condition." Golden-section search, however, appears to be less efficient than "finding a root of the first-order condition using bisection method," because it gets rid of about one-third of the interval in each iteration, as opposed to half. The golden section search provides two additional benefits instead: (i) It does not require $F(x)$ to be differentiable (it has to be differentiable to be able to take the first-order condition), and (ii) in every step, the procedure intends to maximize and not minimize (one cannot tell the difference from just looking at the first-order condition). Thus, using the golden-section search still have some advantages, especially when we don't know the features of $F(x)$ very well.

Example 1: Root finding and optimization

Consider the following static labor-leisure choice problem. A consumer faces a problem

$$\max_{c,\ell} u(c,\ell)$$

subject to

$$c = f(\ell) + x,$$

where c is consumption, ℓ is labor supply, $u(\cdot,\cdot)$ is the utility function, $f(\cdot)$ is the production function, and x is other income.

Assume that the utility function is

$$u(c,\ell) = \ln(c) - \frac{\omega}{\gamma}\ell^{\gamma},$$

where $\omega > 0$ and $\gamma > 0$, and the production function is

$$f(\ell) = \ell^{\alpha},$$

where $\alpha \in (0,1)$. With this specification, the problem is to maximize

$$\ln(\ell^{\alpha} + x) - \frac{\omega}{\gamma}\ell^{\gamma} \tag{10.4}$$

or solve the first-order condition

$$\frac{\alpha\ell^{\alpha-1}}{\ell^{\alpha} + x} - \omega\ell^{\gamma-1} = 0. \tag{10.5}$$

Let $\alpha = 1/3$, $x = 0.5$, $\omega = 1$, and $\gamma = 2$. There are three accompanying MATLAB codes. `Ex1_bisection.m` solves the first-order condition (10.5) numerically using the bisection method to find the optimal $\ell \in [0,1]$. `Ex1_newton.m` solves the first-order condition (10.5) numerically using the Newton-Raphson method to find the optimal ℓ. The code for directly maximizing (10.4) with the Newton method would essentially be the same. Finally, `Ex1_GS.m` solves the optimization problem of maximizing (10.4) using the golden section search.

10.4 Discretizing an AR(1) process

An AR(1) process is a common way to describe exogenous shock process. Computationally, we often wish the represent such a process with a Markov chain that has a similar persistence and volatility. Constructing a Markov chain with similar properties to an AR(1) is called "discretizing" the AR(1). We introduce two methods of discretizing the process

$$z_{t+1} = \rho z_t + \varepsilon_{t+1},$$

where $\rho \in (0, 1)$. The shock ε_{t+1} has the properties $\mathbb{E}_t[\varepsilon_{t+1}] = 0$, $\text{Var}[\varepsilon_{t+1}] = \sigma_\varepsilon^2$, and $\Pr[\varepsilon \leq u] = F(u/\sigma_\varepsilon)$. That is, $F(\cdot)$ is the normalized distribution function for ε_{t+1}. Note this specification implies the unconditional variance of z_t is

$$\sigma_z^2 = \frac{1}{1-\rho^2}\sigma_\varepsilon^2.$$

Our discrete approximation will involve a set of grid points and a transition matrix between them. Let transition matrix have elements π_{ij}, which represents the probability of moving from state i to state j.

10.4.1 Tauchen method

Tauchen (1986) suggests the following method:

1. Create equally spaced grid points, $\{\bar{z}^1, \bar{z}^2, \ldots, \bar{z}^N\}$. Let the distance between points be ω. A common recommendation is to set $\bar{z}^1 = -m\sigma_z$ and $\bar{z}^N = m\sigma_z$, where m represents how wide the coverage of the approximated process is. When $m = 3$, it covers three standard deviations of the original z_t.

2. Set the probabilities

$$\pi_{ij} = \begin{cases} F\left(\dfrac{\bar{z}^1 - \rho\bar{z}^i + \omega/2}{\sigma_\varepsilon}\right) & \text{for } j = 1 \\ F\left(\dfrac{\bar{z}^j - \rho\bar{z}^i - \omega/2}{\sigma_\varepsilon}\right) - F\left(\dfrac{\bar{z}^j - \rho\bar{z}^i + \omega/2}{\sigma_\varepsilon}\right) & \text{for } j = 2, \ldots, N-1 \\ 1 - F\left(\dfrac{\bar{z}^j - \rho\bar{z}^i - \omega/2}{\sigma_\varepsilon}\right) & \text{for } j = N, \end{cases}$$

This approximation is intuitive: for a given $i \in \{1, \cdots N\}$ and $j \in \{2, \cdots, N-1\}$, we suppose we begin at z_i and we compute the probability that the AR(1) assigns to the interval $[\bar{z}^j - \omega/2, \bar{z}^j + \omega/2]$. We then assign this probability to the transition π_{ij}. The end points $j \in \{1, N\}$ follow a similar logic but account for the fact that the support of z extends beyond the grid.

10.4.2 Rouwenhorst method

Some researchers have argued the Tauchen approximation is not accurate when the persistence parameter ρ is close to 1. This issue may be problematic because many stochastic processes used in macroeconomics (e.g., the individual income process, the aggregate productivity process, the firm productivity process) are often very persistent. Some researchers argue that the following method, originally proposed by Rouwenhorst (1995), has better approximation properties.

Below, we introduce a special case of Rouwenhorst's (1995) approximation.[3] The first step is the same as the Tauchen method: Create equally spaced grids, $\{\bar{z}^1, \bar{z}^2, \ldots, \bar{z}^N\}$. Note, though, that the upper bound and lower bound have to be set by $\bar{z}^1 = -m\sigma_z$ and $\bar{z}^N = m\sigma_z$, and m is not arbitrary (see below). Then, we construct the Markov matrix Π_N recursively as follows.

1. For $N = 2$,
$$\Pi_2 = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}.$$

2. For $N \geq 3$, first construct the $N \times N$ matrix
$$\Pi_N = p \begin{bmatrix} \Pi_{N-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix} + (1-p) \begin{bmatrix} \mathbf{0} & \Pi_{N-1} \\ 0 & \mathbf{0}' \end{bmatrix} + p \begin{bmatrix} \mathbf{0}' & 0 \\ \Pi_{N-1} & \mathbf{0} \end{bmatrix} + (1-p) \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \Pi_{N-1} \end{bmatrix},$$
where $\mathbf{0}$ is an $(N-1) \times 1$ zero matrix (column vector). Then, divide all but the top and the bottom rows by two so that the elements in each row are equal to one.

We can set p and m by $p = (1+\rho)/2$ and $m = \sqrt{N-1}$. Then, we can show that the unconditional mean and unconditional variance of the Markov chain are the same as those of the original AR(1) process. This method only uses the information of mean and variance, and the invariant distribution of the Markov process converges to a normal distribution as $N \to \infty$. Thus, the Rouwenhorst method is particularly recommended when ε_{t+1} follows a normal distribution. As was mentioned above, this method is particularly powerful when ρ is close to one.

Example 2: Approximating an AR(1) process

Consider the following AR(1) process:
$$y_{t+1} = 0.95 y_t + \varepsilon_{t+1},$$
where ε_{t+1} follows a normal distribution $N(0, \sigma^2)$ with $\sigma = 0.2$. The code `Ex2_tauchen.m` approximates this process by a 10-state first-order Markov process using the Tauchen method. `Ex2_rouwenhorst.m` performs the same task using the Rouwenhorst method.

[3] See Kopecky and Suen (2010) for further expositions and characterizations.

10.5 Solving a dynamic programming problem with value function iteration

Many different methods are available for solving a dynamic programming problem and we will cover just one of them: value function iteration. This method is particularly robust because it relies on the contraction mapping theorem (see Chapter 4). Consider the Bellman equation of the form

$$V(k) = \max_{k'} \; F(k, k') + \beta V(k'), \tag{10.6}$$

where k is the state variable, k' is the next period value of k, $F(k, k')$ is the return function for the current period, and $\beta \in (0, 1)$ is the discount factor. Now consider the mapping T defined by:

$$TV_i(k) = \max_{k'} \; F(k, k') + \beta V_i(k'). \tag{10.7}$$

Given a function $V_i(k)$ the right-hand side of the equation above defines a new function, which we denote $TV_i(k)$. Now consider starting from $i = 0$ with some guess $V_0(k)$ and iterating $V_{i+1}(k) = TV_i(k)$. The contraction mapping theorem guarantees $V_i(k)$ approaches the true value function $V(k)$ for (10.6) as $i \to \infty$, regardless of the starting value function $V_0(k)$. This theorem suggests the following algorithm for finding $V(k)$:

1. Fix a grid on values of k.
2. Start from an arbitrary $V_0(k)$ represented as the function values on the grid for k.
3. For each k in the grid, solve the right-hand side of (10.7) with $V_0(k)$, that is, $\max'_k \; F(k, k') + \beta V_0(k')$. This step could involve interpolating the values of $V_0(k')$ off of the grid points. Let $V_1(k)$ by this maximized value for each k.
4. Now use $V_1(k)$ on the right-hand side and create $V_2(k)$ as in the previous step. Repeat for $i = 0, 1, 2, ...$ until $V_i(k)$ is similar to $V_{i+1}(k)$.

This algorithm is very robust in the sense that it is based on the theorem that says we will always get to the solution. The downside is that this algorithm can be slow, especially when β is close to one. Many other (potentially faster) algorithms exist, which we don't cover here. Below, we go over this method for several concrete examples.

10.5.1 Deterministic case

Consider the following problem. An infinitely-lived consumer receives a whole cake, whose quantity is a_0, at period 0. The cake can last forever, and the consumer chooses how to eat the cake over time. The consumer's preferences are

$$\sum_{t=0}^{\infty} \beta^t \sqrt{c_t},$$

and the constraint is

$$a_{t+1} = a_t - c_t,$$

where c_t is the consumption of the cake at period t, and a_t is the amount of leftover cake at the beginning of period t. The parameter $\beta \in (0,1)$ is a discount factor. We can write the consumer's utility maximization problem as a Bellman equation:

$$V(a) = \max_{a'} \sqrt{a - a'} + \beta V(a').$$

The algorithm for solving this problem is as follows:

1. Create a grids for values of a. In this case, the natural lower bound is 0, and the natural upper bound is a_0. Call the resulting vector $\begin{bmatrix} a^1 & a^2 & \cdots & a^N \end{bmatrix}'$, where $a^1 = 0$ and $a^N = a_0$. N is the number of grid points.

2. Create the initial value function $V_0(a)$ on the grid points, that is, the vector of $V_0(a^j)$ $(j = 1, 2, ..., N)$. Because the initial value function is arbitrary, we start from the zero vector: $\begin{bmatrix} V_0(a^1) & V_0(a^2) & \cdots & V_0(a^N) \end{bmatrix}' = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}'$.

3. Using the value function $V_0(a)$ created above, for each grid point a^j, solve the right-hand side of the Bellman equation: $\max_{a'} \sqrt{a^j - a'} + \beta V_0(a')$. One way to solve the maximization problem is to restrict a' on the grid points. Then, because we already know the values of $V_0(a')$ on the grid points for a', one can choose a^n that gives the maximum value of $\sqrt{a_j - a^n} + \beta V_0(a^n)$ (with the restriction that $a^n \leq a^j$). An alternative (and better) way is not to restrict the choice of a' to the grid points. Outside the grid points, we can evaluate $V_0(a')$ using the approximation methods we learned in Section 10.1. If $V_0(a')$ is approximated by a differentiable function, we can use the first-order condition and solve the root using the methods in Section 10.2 or we can use the optimization methods in Section 10.3. Note the properties of a contraction mapping ensures the value function is concave in this particular case, and thus, we can use the first-order condition for optimization. When concavity is not obvious, grid search can be a slower but safer optimization method. Store the maximized value as $V_1(a^i)$.

4. Now use $\begin{bmatrix} V_1(a^1) & V_1(a^2) & \cdots & V_1(a^N) \end{bmatrix}'$ on the right-hand side of the Bellman equation, perform the optimization, and create $V_2(a^i)$. In general, one can use the vector of $V_i(a^j)$ to create the vector of $V_{i+1}(a^j)$. Repeat until the vector $V_i(a^j)$ is similar to the vector $V_{i+1}(a^j)$. More concretely, stop when $\max_j |V_{i+1}(a^j) - V_i(a^j)| < \varepsilon$ for a small ε.

Example 3: Deterministic dynamic programming

Consider the above cake-eating problem, with $\beta = 0.98$ and $a_0 = 10$. The provided codes conduct the value function iteration described above and simulate the first 100 periods to plot (i) the time series of the leftover cake a_t, (ii) the time series of consumption c_t, (iii) the value function $V(a)$, and (iv) policy function $a'(a)$. The code `Ex3_gridsearch.m` uses a simple grid search method by restricting the choice a' to be on the grid points for a. The code `Ex3_GS.m` allows the choice for a' to be off the grid points (using the

golden section search for optimization) and interpolate the right-hand side values of the Bellman equation for each potential value for a'.

10.5.2 Stochastic case

Consider a variation of the cake-eating problem. Assume that there is uncertainty: at the beginning of every period t, an additional cake z arrives. The amount of z is stochastic and can be z_H ("high") or z_L ("low": $z_L < z_H$), and the state z, $z \in \{z_H, z_L\}$ follows a Markov process. The probability of the next-period state is z' given the current-period state is z is $\pi_{zz'}$. Now, the consumer maximizes the expected utility

$$\mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t \sqrt{c_t} \right],$$

where $\mathbb{E}_0[\cdot]$ represents the expectation at period 0, and the constraint is now

$$a_{t+1} = a_t - c_t + z_t.$$

The Bellman equation is

$$V(a, z) = \max_{a'} \sqrt{a - a' + z} + \beta \mathbb{E} \left[V(a', z') | z \right],$$

where $\mathbb{E}\left[\cdot|z\right]$ is the conditional expectation given the today's state z. Given the probability structure, we can rewrite it as

$$V(a, z) = \max_{a'} \sqrt{a - a' + z} + \beta \left[\pi_{zH} V(a', H) + (1 - \pi_{zH}) V(a', L) \right].$$

We can still apply the same procedure as the deterministic case. The only difference is that (i) the natural upper bound of the grids is not a_0, given that now a_t can potentially increase; (ii) instead of the vector $V_i(a^j)$, we need to work with two vectors $V_i(a_j, H)$ and $V_i(a_j, L)$ (or a matrix $V_i(a_j, z)$, where $z = H, L$).

As another example, let us consider a version of the stochastic neoclassical growth model introduced in Chapter 7.3. An important difference here is that we allow for elastic labor supply. The representative consumer's utility function is

$$\mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t \left((1 - \phi) \log(C_t) + \phi \log(1 - H_t) \right) \right] \tag{10.8}$$

and the resource constraint is

$$K_{t+1} + C_t = \exp(z_t) K_t^{\alpha} H_t^{1-\alpha} + (1 - \delta) K_t. \tag{10.9}$$

Here, K_t is capital (which is the predetermined variable), C_t is consumption (non-predetermined), and $H_t \in [0, 1)$ is labor supply (non-predetermined). z_t follows the stochastic process

$$z_{t+1} = \rho z_t + \varepsilon_{t+1}, \tag{10.10}$$

where ε_{t+1} is an i.i.d. random variable with mean zero and standard deviation σ. Here, $\beta \in (0,1)$, $\phi \in (0,1)$, $\alpha \in (0,1)$, $\delta \in (0,1)$, $\rho \in (0,1)$, and $\sigma > 0$ are parameters. We consider the social planner's problem of maximizing utility subject to the resource constraint.[4]

In the recursive formulation, the planner's problem can be written as

$$V(K,z) = \max_{K',H} (1-\phi)\log\left(\exp(z)K^{\alpha}H^{1-\alpha} + (1-\delta)K - K'\right) + \phi\log(1-H) + \beta\mathbb{E}\left[V(K',z')|z\right]. \tag{10.11}$$

This equation is very similar to the above cake-eating problem. Only two points are different. First, because the domain of z is now a real line, one has to create discrete grid points, and the stochastic process (10.10) has to be discretized using (for example) the methods introduced above. Second, the planner chooses H.

We can first solve

$$\max_{H} (1-\phi)\log\left(\exp(z)K^{\alpha}H^{1-\alpha} + (1-\delta)K - K'\right) + \phi\log(1-H) \tag{10.12}$$

for a given (z, K, K'). The solution to this problem can be written as $H(z, K, K')$. Note $H(z, K, K')$ may not be an analytically explicit function—we only need the numerical value of H corresponding to a given (z, K, K'). Then, plug in this function to the right-hand side, giving us

$$\begin{aligned} V(K,z) = \max_{K'} \quad & (1-\phi)\log\left(\exp(z)K^{\alpha}H(z,K,K')^{1-\alpha} + (1-\delta)K - K'\right) \\ & + \phi\log(1-H(z,K,K')) + \beta\mathbb{E}\left[V(K',z')|z\right]. \end{aligned}$$

Now, we can use the same procedure as the cake-eating problem.

Example 4: Stochastic neoclassical growth model

Consider the Brock-Mirman model (10.11). Let the stochastic process for $Z_t = \exp(z_t)$ be a two-state Markov process, with two values $H = 1.015$ and $L = 0.985$. The transition probability from i to j $(i, j = L, H)$ is π_{ij}. The parameter values are as follows: $\beta = 0.99$, $\pi_{HH} = \pi_{LL} = 0.95$, $\pi_{LH} = \pi_{HL} = 0.05$, $\alpha = 0.36$, and $\delta = 0.025$. The value of ϕ is set so that the steady-state value of H_t is 1/3. The file `Ex4_gridsearch.m` restricts the choice of K' to be on the grid and uses grid search in the optimization of K'. The solution of H in (10.12) is obtained using the bisection method. After solving the Bellman equation, the program simulates the time series of Y_t, C_t, H_t, $I_t = K_{t+1} - (1-\delta)K_t$, and Z_t, for 3000 periods. After eliminating the first 100 periods, the code logs and HP-filters (using the subroutine `hpfilt.m`; see Chapter 14) each time series and compute the standard deviation and the correlation with Y_t for the cyclical component of each variable. The file `Ex4_GS.m` repeats the same exercise (using the subroutines `Ex4_location.m` and `Ex4_labor.m`), allowing the choice of K' to be outside the grid points using the linear interpolation and the golden section search.

[4] Because of the first and the second welfare theorem, the social planner's solution corresponds to the equilibrium outcome under perfect competition (when everyone has the same wealth).

10.6 Solving a dynamic model with linearization

The previous section solved the dynamic programming problem on a grid of state variables. One would ideally construct the grid to cover the entire portion of the state space where the economy travels.[5] Solution methods of this type are often called "global methods" as they seek to approximate the solution at all relevant state variables. "Local methods" are a different approach in which we choose one point in the state space and examine how the solution changes locally around that point. We often take the deterministic steady state of the economy as our starting point and we can then ask, for example, how consumption changes following a small change in the capital stock. The approach we use is an example a perturbation method, which can approximate the local behavior of the system using simple functions, typically polynomials. In macroeconomics, various degrees of polynomials are used, but in this book, we will only cover the linear (or log-linear) approximation.

These methods are effective to the extent that the linear approximation is accurate. Thus, this method is mainly used for small deviations from the steady state. The deviation can be caused by various reasons, but below we consider a situation where the deviation is due to a random AR(1) shock. In the stochastic neoclassical growth model, for example, we know that capital and consumption converge to their steady-state values when there are no shocks. Thus, it is reasonable to analyze the local dynamics around the steady state when the economy is subject to shocks. In particular, when the shocks are small, a linear solution can yield a reasonable approximation of the true solution.

The benefit of using a linearized system is that it is very fast, especially when the model involves many variables. Directly solving a dynamic programming problem, as in Example 4, is subject to the "curse of dimensionality"—when the number of variables, especially the state variables, is large, solving the problem becomes computationally burdensome very quickly. The solution methods of a large linear system using matrix algebra are well-established, and one can obtain the solution relatively quickly. In addition, because the expectations operator has the affine property, a linear solution has the "certainty equivalence" property. That is, we can solve the model with uncertainty as if it were a deterministic system with corresponding expected values. The dynamics of the system are effectively the same whether there are uncertain shocks or the shock variables move around deterministically.

Below, we explain the principles of solving a linear (or linearized/log-linearized) rational expectations model. Versions of the methods we discuss below (and more advanced procedures) are incorporated into popular software packages such as Dynare (`https://www.dynare.org/`) and are widely used.

When the model is already linear, we can use the methods below directly. When the model is nonlinear, we first linearize or log-linearize the model. We will employ log-linearization below. We have already introduced the idea of log-linearization in Chapter 3. As a result of log-linearization, the equilibrium can be expressed as a system of linear equations with variables that represent log deviations from the steady-state values. Below, suppose we have a system of linear equations that are obtained from such procedures.

Here, we distinguish between two types of variables: predetermined variables and non-

[5] This may not always be possible. For example, if an exogenous state variable has a Gaussian distribution, there could be a realization of this variable that is very extreme. However, one would normally choose the grid so that the economy's state variables are inside the grid with high probability.

predetermined (jump) variables. Predetermined variables are variables whose values are already determined when entering period t. An example is capital stock in the neoclassical growth model. Non-predetermined variables are variables whose values are determined within the period. An example is consumption in the neoclassical growth model.

What do we mean by "solving the model"? Here, our goal is to characterize the movement of endogenous variables as an explicit function of the predetermined variables and exogenous variables. As an example, let us consider the stochastic neoclassical growth model introduced earlier. Recall that the representative consumer's utility is given by (10.8) and the constraint is (10.9). The shocks are given by (10.10). Our eventual goal is to obtain the solution of this model, that is, representing K_{t+1}, C_t, H_t as functions of K_t and z_t.

The first-order conditions lead to the Euler equation

$$\frac{1}{C_t} = \mathbb{E}_t \left[\beta(1 + \alpha \exp(z_{t+1}) K_{t+1}^{\alpha-1} H_{t+1}^{1-\alpha} - \delta) \frac{1}{C_{t+1}} \right] \tag{10.13}$$

and the labor-supply condition

$$\frac{1-\phi}{C_t}(1-\alpha)\exp(z_t)K_t^{\alpha}H_t^{-\alpha} = \frac{\phi}{1-H_t}. \tag{10.14}$$

Log-linearizing (10.9), (10.13), and (10.14), we obtain

$$\bar{K}k_{t+1} + \bar{C}c_t = \bar{K}^{\alpha}\bar{H}^{1-\alpha}(z_t + \alpha k_t + (1-\alpha)h_t) + (1-\delta)\bar{K}k_t,$$

$$-c_t = \mathbb{E}_t[\beta\alpha\bar{K}^{\alpha-1}\bar{H}^{1-\alpha}(z_{t+1} + (\alpha-1)k_{t+1} + (1-\alpha)h_{t+1}) - c_{t+1}],$$

and

$$h_t = \theta(z_t + \alpha k_t - c_t), \tag{10.15}$$

where

$$\theta \equiv \frac{1-\bar{H}}{\bar{H}(1-\alpha)+\alpha}.$$

Here, the lower-case letter is the log deviation from the steady state, that is,

$$x_t \equiv \log\left(\frac{X_t}{\bar{X}}\right)$$

for a variable X_t, where $\bar{X}$ is the steady-state value.[6] For future use, let us use (10.15) to eliminate h_t from the first two equations. Rearranging, we obtain

$$\bar{K}k_{t+1} = \bar{K}^{\alpha}\bar{H}^{1-\alpha}(1+\theta(1-\alpha))z_t + (\bar{K}^{\alpha}\bar{H}^{1-\alpha}\alpha(1+\theta(1-\alpha)) + (1-\delta)\bar{K})k_t - (\bar{K}^{\alpha}\bar{H}^{1-\alpha}\theta(1-\alpha) + \bar{C})c_t, \tag{10.16}$$

and (using $\mathbb{E}_t[z_{t+1}] = \rho z_t$ from (10.10))

$$\begin{aligned}\beta\alpha\bar{K}^{\alpha-1}\bar{H}^{1-\alpha}(1-\alpha)(1-\theta\alpha)k_{t+1} + (\beta\alpha\theta(1-\alpha)\bar{K}^{\alpha-1}\bar{H}^{1-\alpha}+1)\mathbb{E}_t[c_{t+1}] \\ = \beta(1+\theta(1-\alpha))\bar{K}^{\alpha-1}\bar{H}^{1-\alpha}\rho z_t + c_t.\end{aligned} \tag{10.17}$$

[6] In log-linearizing (10.13), we applied the log-liniarization technique ignoring the expectation parameter on the right-hand side. This step exploits the certainty equivalence property discussed in Chapter 7.3.4.

10.6.1 Blanchard-Kahn condition

Blanchard and Kahn (1980) derive a condition for the uniqueness of the equilibrium in linear rational expectation models. This condition, called the Blanchard-Kahn condition, also helps solve the model. Before introducing the general result, let us fix the idea with a few examples.

One non-predetermined variable

Consider the dynamic system with one non-predetermined variable y_t:

$$y_t = \phi y_{t+1}, \tag{10.18}$$

where $\phi > 0$ is a parameter. For convenience, let us rewrite this relationship

$$y_{t+1} = \lambda y_t, \tag{10.19}$$

where $\lambda = 1/\phi$. The reason we wrote (10.18) first is to emphasize this relationship is not to imply y_{t+1} is predetermined by y_t. One can imagine the relationship between consumption today and consumption tomorrow in the consumer's Euler equation.

Suppose that one of the equilibrium conditions is that y_t does not "blow up," that is, $\lim_{T\to\infty} y_{t+T}$ remains finite. Two questions can be asked here. First, what condition on ϕ would guarantee the uniqueness of the equilibrium? Second, under that condition, what are the properties of the equilibrium?

Applying (10.19) repeatedly, we obtain

$$y_{t+T} = \lambda^T y_t.$$

When $\lambda > 1$, the only possible value of y_t for $\lim_{T\to\infty} y_{t+T}$ remaining finite is $y_t = 0$. Thus, the equilibrium is unique. When $\lambda \leq 1$, any value of y_t is consistent with $\lim_{T\to\infty} y_{t+T}$ being finite. This situation is often called "indeterminacy." Thus, the condition for the uniqueness is $\lambda > 1$, and the solution is

$$y_t = 0 \text{ for all } t. \tag{10.20}$$

Once again, the particular logic here is important: to guarantee the uniqueness of the equilibrium, λ has to be such that if we don't choose the right y_t, the economy will eventually "blow up" meaning the variables will diverge. Then, the only "right" y_t is the unique solution, which is (10.20) in this case.

One non-predetermined variable and one predetermined (exogenous) variable

Now, assume y_t is stochastic due to the influence of the exogenous (and predetermined) shock z_t. Consider the dynamic system

$$y_t = \phi \mathbb{E}_t[y_{t+1}] + \hat{z}_t,$$

where $\phi > 0$ is a parameter, $\hat{z}_t$ is a shock that follows

$$\hat{z}_{t+1} = \rho \hat{z}_t + \hat{\varepsilon}_{t+1}$$

with $\rho \in (0,1)$, and $\hat{\varepsilon}_{t+1}$ is an i.i.d. random variable whose mean is 0 and the standard deviation is $\hat{\sigma} > 0$. Here, $\hat{z}_t$ is predetermined; that is, the value of $\hat{z}_t$ is determined at the value of $\hat{z}_{t-1}$ and a shock $\hat{\varepsilon}_t$, which is revealed at the beginning of period t.

Again, we rewrite

$$\mathbb{E}_t[y_{t+1}] = \lambda y_t + z_t, \tag{10.21}$$

where $\lambda = 1/\phi$ and $z_t = -\hat{z}_t/\phi$. Therefore,

$$z_{t+1} = \rho z_t + \varepsilon_{t+1}, \tag{10.22}$$

where ε has mean zero and standard deviation $\sigma = \hat{\sigma}/\phi$.

Let us look for the solutions where $\lim_{T\to\infty} E_t[y_{t+T}]$ is finite. The relationship (10.21) implies

$$\mathbb{E}_{t+1}[y_{t+2}] = \lambda y_{t+1} + z_{t+1},$$

and thus,

$$\mathbb{E}_t[y_{t+2}] = \mathbb{E}_t[\mathbb{E}_{t+1}[y_{t+2}]] = \lambda \mathbb{E}_t[y_{t+1}] + \mathbb{E}_t[z_{t+1}] = \lambda^2 y_t + \lambda z_t + \rho z_t = \lambda^2 \left[y_t + \frac{1}{\lambda}\left[1 + \frac{\rho}{\lambda}\right] z_t \right],$$

where the first equality utilizes the law of iterated expectations. Repeating this procedure (assuming $\lambda \neq \rho$),

$$\mathbb{E}_t[y_{t+T}] = \lambda^T \left[y_t + \frac{1}{\lambda}\left[1 + \frac{\rho}{\lambda} + \left(\frac{\rho}{\lambda}\right)^2 + \cdots + \left(\frac{\rho}{\lambda}\right)^{T-1}\right] z_t \right] = \lambda^T \left[y_t + \frac{1-(\rho/\lambda)^T}{\lambda - \rho} z_t \right].$$

Thus, when $\lambda > 1$, the only situation where $\lim_{T\to\infty} E_t[y_{t+T}]$ is finite is when

$$y_t = -\lim_{T\to\infty} \frac{1-(\rho/\lambda)^T}{\lambda - \rho} z_t = -\frac{1}{\lambda - \rho} z_t, \tag{10.23}$$

and this expression gives the unique solution of y_t. When $\lambda \leq 1$, indeterminacy exists, because many values of y_t can keep $\lim_{T\to\infty} E_t[y_{t+T}]$ finite.[7]

The "Guess and verify" method (the method of undetermined coefficients)

If we *know* the solution is unique, we can use the "guess and verify" (the method of undetermined coefficients) to find the solution, similarly to Section 4.4.4. Suppose ("guess") that the solution of (10.21) takes the form

$$y_t = \mathcal{A} z_t, \tag{10.24}$$

where $\mathcal{A}$ is an unknown constant. Plugging this equation into (10.21), we obtain

$$\mathbb{E}_t[\mathcal{A} z_{t+1}] = \lambda \mathcal{A} z_t + z_t. \tag{10.25}$$

[7] In the above expression for $\mathbb{E}_t[y_{t+T}]$,

$$\lambda^T \left[y_t + \frac{1-(\rho/\lambda)^T}{\lambda - \rho} z_t \right] = \lambda^T y_t + \frac{\lambda^T - \rho^T}{\lambda - \rho} z_t,$$

which would remain finite when $\lambda \leq 1$ (note we assumed $\rho < 1$).

From (10.22), the left-hand side of (10.25) is equal to $\mathcal{A}\rho z_t$, and thus we can rearrange (10.25) as

$$(\mathcal{A}\rho - \mathcal{A}\lambda - 1)z_t = 0.$$

This equality has to hold for all z_t; thus, $\mathcal{A}\rho - \mathcal{A}\lambda - 1$ must be zero. Therefore,

$$\mathcal{A} = -\frac{1}{\lambda - \rho},$$

which, in the equation (10.24), yields the same solution as (10.23). This solution indeed satisfies (10.21) ("verify").

m non-predetermined variables and $n + k$ predetermined variables

Now, let us consider a more general situation. Suppose the model now has m non-predetermined variables and $n + k$ predetermined variables. Predetermined variables include n endogenous variables, such as K_t in the growth model example, and k exogenous variables, such as z_t in the growth model. Suppose the model can be expressed as the system of equations

$$B \begin{bmatrix} x_{t+1} \\ \mathbb{E}_t[y_{t+1}] \end{bmatrix} = A \begin{bmatrix} x_t \\ y_t \end{bmatrix} + Ea_t, \tag{10.26}$$

where x_t is an $n \times 1$ vector of endogenous predetermined variables, y_t is an $m \times 1$ vector of non-predetermined variables, and B and A are $(n + m) \times (n + m)$ matrices.[8] The vector a_t represents exogenous predetermined variables, which is a $k \times 1$ vector. E is an $(n + m) \times k$ matrix. Below, we consider a situation where each element of a_t follows the AR(1) process

$$a^i_{t+1} = \rho a^i_t + \varepsilon^i_{t+1},$$

where a^i_t is the element i of a_t vector, $\rho \in [0, 1)$ is the common persistence parameter, and ε_{t+1} is the mean zero i.i.d. shock for element i.[9]

In our stochastic growth example, $x_t = k_t$, $y_t = c_t$, and $a_t = z_t$. In the matrix form, (10.16) and (10.17) can be expressed as

$$B \begin{bmatrix} k_{t+1} \\ \mathbb{E}_t[c_{t+1}] \end{bmatrix} = A \begin{bmatrix} k_t \\ c_t \end{bmatrix} + Ez_t, \tag{10.27}$$

where

$$B = \begin{bmatrix} \bar{K} & 0 \\ \beta\alpha\bar{K}^{\alpha-1}\bar{H}^{1-\alpha}(1-\alpha)(1-\theta\alpha) & \beta\alpha\theta(1-\alpha)\bar{K}^{\alpha-1}\bar{H}^{1-\alpha} + 1 \end{bmatrix},$$

$$A = \begin{bmatrix} \bar{K}^{\alpha}\bar{H}^{1-\alpha}\alpha(1+\theta(1-\alpha)) + (1-\delta)\bar{K} & -(\bar{K}^{\alpha}\bar{H}^{1-\alpha}\theta(1-\alpha) + \bar{C}) \\ 0 & 1 \end{bmatrix},$$

and

$$E = \begin{bmatrix} \bar{K}^{\alpha}\bar{H}^{1-\alpha}(1+\theta(1-\alpha)) \\ \beta(1+\theta(1-\alpha))\bar{K}^{\alpha-1}\bar{H}^{1-\alpha}\rho \end{bmatrix}.$$

[8] Some models cannot be expressed in this manner, but a broad class of macroeconomic models can be transformed to fit into this expression (after linearizing or log-linarizing).

[9] A common ρ is assumed here for the ease of exposition.

Suppose B is invertible in (10.26).[10] Then we can rewrite it as

$$\begin{bmatrix} x_{t+1} \\ \mathbb{E}_t[y_{t+1}] \end{bmatrix} = F \begin{bmatrix} x_t \\ y_t \end{bmatrix} + Ga_t, \tag{10.28}$$

where $F = B^{-1}A$ and $G = B^{-1}E$. The matrix F can be decomposed into (Jordan decomposition)

$$F = HJH^{-1}. \tag{10.29}$$

The matrix J is a diagonal matrix with eigenvalues of F on the diagonal, and H is a matrix that consists of the corresponding eigenvectors:

$$J = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n+m} \end{bmatrix}$$

and

$$H = \begin{bmatrix} v_1 & v_2 & \cdots & v_{n+m} \end{bmatrix}.$$

(For exposition, we are limiting ourselves to the situation where eigenvalues are real and distinct.) Here, we line up the eigenvalues with the ascending order of the absolute value: $|\lambda_1| < |\lambda_2| < \cdots < |\lambda_{n+m}|$. Let h denote the number of eigenvalues that satisfies $|\lambda_i| > 1$. Blanchard and Kahn (1980) show that the equilibrium (i.e., the solutions for x_t and y_t that don't blow up in the future) is unique if $h = m$. That is, the number of eigenvalues that are outside the unit circle is the same as the number of non-predetermined variables. This condition is often called the Blanchard-Kahn condition.

The intuition is the same as in the earlier case with one non-predetermined variable. Here, we deal with the matrix F. In general, a matrix multiplication to a vector is a combination of expanding (or contracting) and rotating the vector. The eigenvalues tell us whether the multiplication expands ($|\lambda_i| > 1$) or contracts ($|\lambda_i| < 1$) the vector in the directions that are indicated by the eigenvectors. When $h = m$, we can choose a unique set of non-predetermined variables such that the vector $\begin{bmatrix} x_t' & y_t' \end{bmatrix}'$ is "zero" in the direction of expansions. The following procedure can identify such $\begin{bmatrix} x_t' & y_t' \end{bmatrix}'$.

Using (10.29) on (10.28) and multiplying H^{-1} from the left,

$$H^{-1} \begin{bmatrix} x_{t+1} \\ \mathbb{E}_t[y_{t+1}] \end{bmatrix} = JH^{-1} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + H^{-1}Ga_t. \tag{10.30}$$

Let us partition H^{-1} and J with n and m rows and n and m columns and call

$$H^{-1} = \begin{bmatrix} \tilde{H}_{11} & \tilde{H}_{12} \\ \tilde{H}_{21} & \tilde{H}_{22} \end{bmatrix}$$

[10] In general, B may not be invertible. For example, in our stochastic growth example, if we did not use (10.15) to eliminate h_t earlier and treated h_t as another non-predetermined variable, the resulting 3×3 matrix B would not have been invertible. In such a case, we can use the Generalized Schur decomposition (also called the QZ decomposition) instead of the Jordan decomposition below. See, for example, Heer and Maußner (2024), Chapters 2 and 3.

and

$$J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}.$$

Note J_1 and J_2 are both diagonal matrices. Let

$$H^{-1}G = \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix},$$

where Γ_1 is an $n \times k$ matrix and Γ_2 is an $m \times k$ matrix. Let

$$\begin{bmatrix} \tilde{x}_t \\ \tilde{y}_t \end{bmatrix} = \begin{bmatrix} \tilde{H}_{11} & \tilde{H}_{12} \\ \tilde{H}_{21} & \tilde{H}_{22} \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix}. \tag{10.31}$$

Then (10.30) can be rewritten as

$$\begin{bmatrix} \tilde{x}_{t+1} \\ \mathbb{E}_t[\tilde{y}_{t+1}] \end{bmatrix} = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix} \begin{bmatrix} \tilde{x}_t \\ \tilde{y}_t \end{bmatrix} + \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix} a_t.$$

Now let us look at the last m elements of the equation:

$$\mathbb{E}_t[\tilde{y}_{t+1}] = J_2 \tilde{y}_t + \Gamma_2 a_t.$$

One can immediately see the similarities to (10.21). In fact, because J_2 is diagonal, we can apply exactly the same steps as before to each element of this set of equations and show that the solution (that guarantees $\mathbb{E}_t[\tilde{y}_{t+T}]$ remains finite as $T \to \infty$) is

$$\tilde{y}_t = \Lambda \Gamma_2 a_t,$$

where Λ is an $m \times m$ diagonal matrix whose diagonal elements are $-1/(\lambda_i - \rho)$ (i runs from $n+1$ to $n+m$). From (10.31), this expression for $\tilde{y}_t$ implies

$$\Lambda \Gamma_2 a_t = \tilde{H}_{21} x_t + \tilde{H}_{22} y_t.$$

Therefore, the solution for y_t is

$$y_t = -\tilde{H}_{22}^{-1} \tilde{H}_{21} x_t + \tilde{H}_{22}^{-1} \Lambda \Gamma_2 a_t.$$

From (10.28),

$$x_{t+1} = F_{11} x_t + F_{12} y_t + G_1 a_t = (F_{11} - F_{12} \tilde{H}_{22}^{-1} \tilde{H}_{21}) x_t + (G_1 + F_{12} \tilde{H}_{22}^{-1} \Lambda \Gamma_2) a_t.$$

Thus, we obtained the solution.

Example 5: Solving the log-linearized stochastic neoclassical growth model

Consider the log-linearized stochastic neoclassical growth model (10.27). The parameter values are the same as Example 4, except that the shock process follows (10.10) with $\rho = 0.95$. The code `Ex5_BK.m` follows the above steps to solve the model. One can see that in the matrix J, $J(1,1) = 0.9537$ and $J(2,2) = 1.0592$, so that the Blanchard-Kahn condition is satisfied. The solution is

$$c_t = 0.5691 k_t + 0.3920 z_t,$$

$$k_{t+1} = 0.9537k_t + 0.1132z_t,$$

and

$$h_t = -0.2431k_t + 0.7070z_t.$$

The code also plots the log-linearized impulse response functions (see Section 3.5.2) and also compute the same HP-filtered statistics as in Example 4, with $\sigma = 0.007$.

Part III

Applications

Chapter 11

Consumption

Giovanni L. Violante

11.1 Introduction

Household consumption is a key determinant of welfare and, as a result, it plays a fundamental role in many areas of macroeconomics, such as growth, business cycles, inequality, taxation, and asset pricing. The growth of per-capita consumption over time is an unequivocal sign of augmented prosperity of a society. Because consumption fluctuations over the business cycle are costly to households, both fiscal and monetary policy go to great lengths in order to stabilize aggregate consumption expenditures. The distribution of consumption in the population is a credible measure of inequality in standard of living across households, more so than income. Crucial objectives of redistributive and social insurance policies are that of supporting a minimum level of consumption above poverty for all households, and that of limiting the pass through of income losses to household spending (and thus, to their well-being). Finally, consumption choices of investors over time and across states determine stochastic discount factors which price assets in financial markets.

In light of this centrality, it is not surprising that the theory and empirics of consumption choices has, historically, attracted so much attention from economists. As Deaton (1992a) puts it, in the preface of his book, *attempts by economists to understand the saving and consumption patterns of households have generated some of the best science in economics.* The desire to microfound the empirical relation between consumption and income, which contradicted the simple Keynesian consumption function where expenditures are modelled as a linear function of current income, sparked some of the first examples of forward-looking dynamic optimization (Modigliani and Brumberg, 1954; Friedman, 1957) and, since then, led to gradual enrichment of the optimization framework. At the same time, the ever wider availability of large microeconomic data sets on income and expenditures (first survey data, now "big" administrative and proprietary data) created fertile grounds for the application of state-of-the art econometric techniques to test models and quantify key magnitudes. In the last 30 years, this theoretical and empirical advances have been incorporated into general equilibrium models with heterogeneous households that constitute one of the main workhorses for the study of business cycle, inequality and government policy in macroeconomics.

This chapter offers an introduction to this vast literature. It emphasizes the theoretical advancements in this field, but it also makes an attempt to relate models to data. The chap-

ter is organized as follows. Section 11.2 derives consumption allocations under two extreme financial market structures: autarky and complete markets. It then argues that the empirical evidence suggests a market structure in between these two which offers "partial insurance" against income shocks, and introduces an economy where only a non-state-contingent bond is traded. Section 11.3 studies in some depth the optimal intertemporal consumption/saving problem of a household who can save and borrow through this asset, the so called "income fluctuation problem". We start from the deterministic case, and then we analyze the stochastic case with random income fluctuations. The permanent income hypothesis, where certainty equivalence holds and risk plays a role for consumption allocation, is a special case of this environment. We then move beyond certainty equivalence and analyze environments where risk matters for consumption and saving. We analyze the two sources of precautionary saving, prudence and occasionally binding borrowing constraints, and explain how they induce concavity in the consumption function. In Section 11.4, we combine a continuum of households facing income fluctuation problems, and study how they give rise to an endogenous joint distribution of income, consumption, and wealth. We then characterize the stationary equilibrium of these economies, first without and then with production. This last section extends the analysis of Section 7.6.

Because of space constraints, we omit developing a number of interesting and important topics related to consumption. For example, business cycles and asset pricing, life-cycle patterns, information frictions, consumer durables and housing, bequest, habits, and non-standard preferences. Some of these topics will be covered in later chapters, such as Chapters 14, 16 and 21. We also refer the reader to the surveys by Hall (1988b), Muellbauer (1994), Browning and Lusardi (1996), Attanasio (1999), Browning and Crossley (2001), Campbell (2003), Attanasio and Weber (2010), Meghir and Pistaferri (2011), Piazzesi and Schneider (2016), Kaplan and Violante (2022), and the thematic books by Deaton (1992a) and Jappelli and Pistaferri (2020) for additional material.

11.2 Consumption under autarky and full insurance

Consider an endowment economy with aggregate uncertainty, as the one outlined in Chapter 7 of the book. Let $\omega_t \in \Omega$ (a finite set) be the realization of a stochastic event (e.g., an aggregate shock) at date t. Let $\omega^t = \{\omega_0, \omega_1, ..., \omega_t\}$ be the history of events until time t, with $\omega^t \in \Omega^t \equiv \Omega \times \Omega \times ... \times \Omega$, the $t+1$ Cartesian product of Ω. Each history ω^t has unconditional probability of occurring $\pi(\omega^t)$. We assume all households have rational expectations, i.e. they forecast by using the true probability distribution. The economy is populated by a continuum of measure one of infinitely-lived households indexed by i who are endowed with stochastic income $y_{i,t}(\omega^t)$ such that

$$\int_0^1 y_{i,t}\left(\omega^t\right) di = Y_t\left(\omega^t\right),$$

where $Y_t(\omega^t)$ is the (random) aggregate endowment of the economy. Each realization $\omega_t \in \Omega$ corresponds to a particular value of the aggregate endowment and a particular distribution

of it across households.[1]

Households are expected utility maximizers, with period utility $u\left(c_{i,t}\left(\omega^t\right)\right)$, where $c_{i,t}\left(\omega^t\right)$ is consumption of individual i upon realization of history ω^t. We assume that u satisfies standard properties, i.e. $u' > 0$ and $u'' < 0$. Let $C_t\left(\omega^t\right) = \int_0^1 c_{i,t}\left(\omega^t\right) di$ denote aggregate consumption. Aggregate feasibility implies that aggregate consumption equals the aggregate endowment along each history

$$C_t\left(\omega^t\right) = Y_t\left(\omega^t\right) \text{ for all } t, \omega^t \in \Omega^t.$$

The individual consumption allocation that arises in the equilibrium of this economy depends on market arrangements. We consider two extreme benchmarks: autarky and full insurance.

Under autarky, there is no insurance market which allows individuals to trade across states, and no storage technology to transfer resources across time (e.g., the endowment is fully perishable). In this economy, an individual i who receives a random stream of income shocks $\left\{\left\{y_{i,t}\left(\omega^t\right)\right\}_{\omega^t \in \Omega^t}\right\}_{t=0}^{\infty}$ has no other choice than consuming their income in every state:

$$c_{i,t}\left(\omega^t\right) = y_{i,t}\left(\omega^t\right), \text{ for all } t, \omega^t \in \Omega^t. \tag{11.1}$$

and equation (11.1) is also their budget constraint. Under autarky there is full pass-through of individual income shocks into consumption, or no consumption smoothing whatsoever.

Consider now the other end of the spectrum of market arrangements: complete markets (also called full insurance or full risk-sharing), introduced earlier in Section 7.4. Under this arrangement, households can trade a complete set of Arrow securities. This market structure allows every individual i to achieve any transfer of income across states and across time, as long as these trades respect the time-zero Arrow-Debreu budget constraint

$$\sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega^t} p_t\left(\omega^t\right)\left[c_{i,t}\left(\omega^t\right) - y_{i,t}\left(\omega^t\right)\right] = 0. \tag{11.2}$$

The Arrow-Debreu competitive equilibrium for this economy was defined in Section 7.4, where it is shown that for any pair of households (i, j),

$$\frac{u'\left(c_{i,t}\left(\omega^t\right)\right)}{u'\left(c_{j,t}\left(\omega^t\right)\right)} = \frac{\lambda_i}{\lambda_j}, \text{ for all } t, \omega^t \in \Omega^t, \text{ and } (i, j), \tag{11.3}$$

with λ_i and λ_j representing Lagrange multipliers on their lifetime budget constraint.

Equation (11.3) illustrates the defining property of consumption allocations under full insurance: when households can trade a complete set of Arrow securities, *the ratio of marginal utility of consumption of any two households is constant across time and states.*

To make further analytical progress, note that in the special case of CRRA utility, where

$$u\left(c_{i,t}\left(\omega_t\right)\right) = \frac{c_{i,t}\left(\omega_t\right)^{1-\sigma}}{1-\sigma}, \text{ with } \sigma \in [0, \infty)$$

[1] For example, with $I = 2$ and $\omega_t \in \left\{\omega^L, \omega^H\right\}$, we could have a configuration where $Y\left(\omega^L\right) = 2, Y\left(\omega^H\right) = 4$, and $y_1\left(\omega^L\right) = 2, y_2\left(\omega^L\right) = 0$, $y_1\left(\omega^H\right) = 1, y_2\left(\omega^H\right) = 3$. Thus, the aggregate endowment in state H is larger than in state L, but type 1 is better off in state L and type 2 in state H.

equation (11.3) becomes

$$\frac{c_{i,t}\left(\omega^t\right)}{c_{j,t}\left(\omega^t\right)} = \left(\frac{\lambda_i}{\lambda_j}\right)^{-\frac{1}{\sigma}},$$

which implies that the ratio of consumption allocations (not just marginal utility of consumption) between households is constant across states and over time. Summing over $i = 1, ..., I$ on both sides yields

$$c_{j,t}\left(\omega^t\right) = \left[\frac{\left(\lambda_j\right)^{-\frac{1}{\sigma}}}{\int_0^1 \left(\lambda_i\right)^{-\frac{1}{\sigma}} di}\right] C_t\left(\omega^t\right), \tag{11.4}$$

so, individual consumption is proportional to aggregate consumption (or aggregate endowment), with a coefficient of proportionality, in the square bracket, that depends on the relative values of λ.[2] Thus, individual consumption is not constant over time, but it is independent of the realization of the individual endowment. Unanticipated shocks to the aggregate endowment are the only source of individual variation in consumption. These fluctuations cannot be insured because, by definition, they are common across households.

11.2.1 Full insurance with preference heterogeneity

How does preference heterogeneity affect consumption allocations under complete markets? It is immediate to see from (11.3) that the hallmark of complete markets —constant ratio of marginal utility of consumption across households— is still valid even with heterogeneity in u. It is no longer necessarily true, however, that the ratio of consumption is also equalized. To see this, assume u is CRRA and households (indexed by i) differ with respect to the curvature parameter σ

$$u^i\left(c_{i,t}\left(\omega_t\right)\right) = \frac{c_{i,t}\left(\omega_t\right)^{1-\sigma_i}}{1-\sigma_i}.$$

Consider an economy with two agents $i = 1, 2$. The individual FOCs conditions combined with the market clearing condition imply

$$\lambda_2 c_{1,t}\left(\omega^t\right)^{-\sigma_1} = \lambda_1\left[Y_t\left(\omega^t\right) - c_{1,t}\left(\omega^t\right)\right]^{-\sigma_2} \tag{11.5}$$

Figure 11.1 plots right-hand side and left-hand side as a function of $c_{1,t}$ for the case where type 1 is less risk averse than type 2 $(\sigma_1 < \sigma_2)$. The two curves cross only once. Now consider a rise in the aggregate endowment. We know from (11.4) that when $\sigma_1 = \sigma_2$ consumption would increase proportionately so to leave the ratio of consumption between the two households unchanged. With unequal risk aversion, instead, it is efficient for the planner to have the consumption of the least risk-averse households (type 1 in our example) fluctuate more in response to aggregate shocks, as evident from Figure 11.1.

11.2.2 Empirical tests of the full insurance hypothesis

Abstracting from heterogeneity in preference for risk, one can combine the consumption allocation under autarky in (11.1) and complete markets in (11.4) to derive an encompassing

[2] We have derived this result in Section 5.2.1. There, our formulation did not include uncertainty and we have derived it directly from the equilibrium conditions.

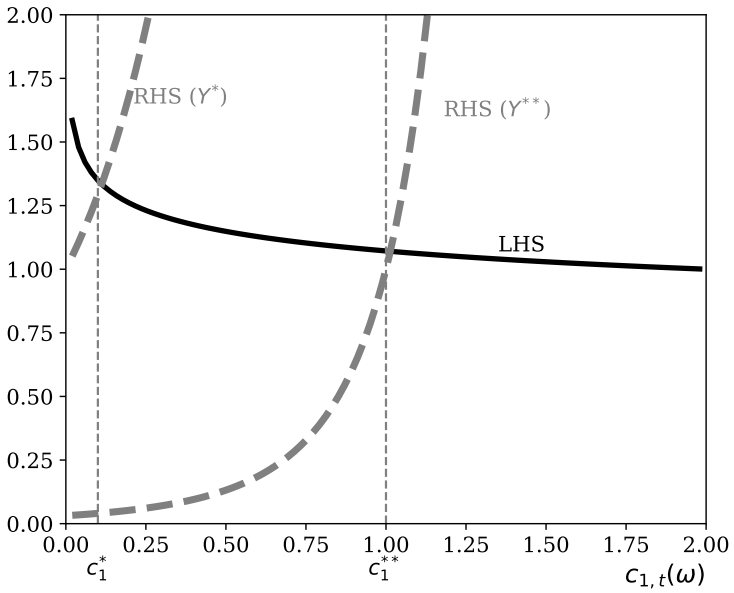


Figure 11.1: Plot of the left-hand-side (LHS) and right-hand-side (RHS) of equation (11.5).

Notes: Parameter values: $\sigma_1 = 0.1, \sigma_2 = 5, \lambda_1 = \lambda_2 = 1$. Aggregate endowment rises from $Y^* = 1$ to $Y^{**} = 2$. Agent 1, who is less risk averse, absorbs most of the change in the endowment.

empirical model for consumption growth that can be taken to the data:

$$\Delta \log c_{i,t} = \beta_1 \Delta \log C_t + \beta_2 \Delta \log y_{i,t} + \varepsilon_{i,t},$$

where $y_{i,t}$ is current individual income, C_t is aggregate consumption and $\varepsilon_{i,t}$ is an error term independent of the two regressors.[3] The autarky hypothesis implies $\beta_1 = 0$ and $\beta_2 = 1$, i.e. individual consumption tracks individual income. The full risk-sharing hypothesis, instead, implies $\beta_1 = 1$ and $\beta_2 = 0$, i.e. individual consumption tracks aggregate endowment, but is independent of individual income.

In the early 1990s' a large empirical literature developed with the aim of testing these hypotheses using longitudinal micro data on consumption expenditures and income (Altug and Miller, 1990, Mace, 1991, Cochrane, 1991, Nelson, 1994, Townsend, 1994, Attanasio and Davis, 1996, Hayashi, Altonji, and Kotlikoff, 1996, and Jappelli and Pistaferri, 2006). Overall, the conclusion of this literature is that an empirically plausible model of consumption behavior lies somewhere in one in between full risk-sharing and autarky: individual income shocks are only *partially insured*.

More details

Mace (1991) used the short panel dimension of the Consumer Expenditure Survey (CES), the main survey on consumption expenditures of U.S. households, to test the relation

[3]This equation can also be interpreted as the result of a log-linearization for more general preferences.

between individual and aggregate consumption growth, and rejected the null hypothesis that $\beta_1 = 1$. Cochrane (1991), instead, tested the null hypothesis that individual consumption does not react to idiosyncratic income fluctuations ($\beta_2 = 0$). He used the Panel Study of Income Dynamics (PSID), the longest-running longitudinal dataset representative of the U.S. population, to identify explicit events associated with income losses, such as days of work lost because of illness, involuntary job loss, weeks spent by jobless household heads looking for employment, days of work lost to strikes. He concluded that household food consumption expenditures (the only spending category well measured in this dataset at the time) are quite —although not fully— responsive to many of these indicators, a finding that contradicts the efficient risk sharing hypothesis. Attanasio and Davis (1996) analyzed the relation between changes in consumption and changes in relative earnings across demographic groups (defined by cohorts and education) in the U.S. in the 1980s. Their results indicate a "spectacular failure" of the full insurance hypothesis across groups.

According to the full risk-sharing hypothesis under CRRA, individual consumption moves in lockstep with aggregate consumption for all households. As a result, the household ranking (i.e., their relative position) in the consumption distribution remains constant over time. The hypothesis of lack of rank mobility in household consumption was tested, and amply rejected, by Fisher and Johnson (2006) and Jappelli and Pistaferri (2006).

Tests of perfect risk sharing were extended to the case of risk aversion heterogeneity by Mazzocco and Saini (2012) and Schulhofer-Wohl (2011). As discussed by these authors, preference heterogeneity can lead to a spurious rejection of this hypothesis. To see this, note that, as explained above, if households differ in their degree of curvature in utility, optimality conditions for the CRRA case imply that

$$\Delta \log c_{i,t} = \beta_{i,1} \Delta \log C_t + \beta_2 \Delta \log y_{i,t} + \varepsilon_{i,t} \tag{11.6}$$

where $\beta_{i,1}$ is larger the lower is risk aversion of household i relative to its population average. Thus, if one estimates the misspecified equation (11.4) instead of (11.6), the error includes the term $(\beta_{i,1} - 1)\Delta \log C_t$. If earnings of low risk aversion (high $\beta_{i,1}$) individuals are more procyclical, then this omitted variable would induce a positive bias on the estimate of β_2. One reasons to expect that less risk-averse households will have more procyclical incomes is that more risk tolerant workers will choose occupations that carry more risk, both idiosyncratic and aggregate. Schulhofer-Wohl (2011) uncovers some evidence about this mechanism from respondents in the Health and Retirement Study.

11.2.3 Two approaches to partial risk sharing

How do we go about developing and quantifying economic theories of partial risk sharing? Economists have followed two methodologies. The first one —which we may call the *endogenous incomplete markets* approach— is rooted in the tradition of microfoundation of macroeconomics. According to this approach, one should model explicitly the fundamental frictions that undermine the emergence of full insurance in the competitive equilibrium.

Recall that a complete set of state-contingent securities will be traded if two assumptions hold: (i) perfect enforcement of contracts and (ii) perfect information. Deviations from these assumptions, such as limited commitment or private information (adverse selection or moral hazard), lead to partial risk sharing. There exist several examples of models that incorporate these frictions in the competitive equilibrium and analyze implications for the distribution of consumption allocations. For example, building on Kehoe and Levine (1993, 2001), Krueger and Perri (2006) relax the perfect contract enforcement assumption; Doepke and Townsend (2006) and Attanasio and Pavoni (2011) relax the perfect information assumption.

The principal strength of this methodology is that the market structure is not assumed exogenously, but it emerges in equilibrium and, as a result, is endogenous with respect to changes in primitives of the economy and to government policy.[4] The main shortcomings of this strategy are two. First, the financial contracts that end up being traded in equilibrium are quite complex, e.g. they are history dependent and state-contingent, and thus very different from the simple ones we observe in reality. For example, in the limited enforcement economy default is an off-equilibrium threat, but it never actually happens, whereas households do default in the real world. Second, the empirical implications of these models for consumption allocations are often rejected by the data. Take, again, the limited commitment model as an example. Consumption always drifts down when the participation constraint of the household does not bind, and it jumps up when it does bind, which happens whenever income increases sufficiently (Kocherlakota, 1996). This pattern gives rise to an extreme and counterfactual degree of left-skewness in the consumption distribution (Broer, 2013). This disconnect between theory and data represents a challenge for quantitative analysis of these models.

This criticism fueled the second methodology, which we can call the *exogenous incomplete markets* approach. This alternative view prescribes that we should only model the assets that we observe in the real world, e.g. non state-contingent bonds, risky publicly traded equity, housing, privately held firms, etc. The advantage of this perspective is that it leads to models that can be easily and naturally taken to the data. Its drawback is that the asset market structure is exogenously assumed and, as a result, it does not respond to changes in model parameters. To be precise, in an environment with exogenous incomplete markets, a shift in primitives (e.g., in the stochastic process of income shocks) does not lead to the addition or the disappearance of certain securities being traded, but it does impact the equilibrium prices at which existing assets are traded (and, possibly, also impacts the value of borrowing limits, depending on how they are specified). As a result, the pass-through of income shocks to consumption, an indicator of the degree of risk sharing, is also affected.

Before articulating the analysis of this approach, it is worth noting that there exist some results in the literature which show that the right combination of fundamental frictions can give rise to a realistic market structure emerging endogenously in equilibrium. Notable examples are Allen (1985) and Cole and Kocherlakota (2001) who consider an environment with unobservable income shocks and hidden saving, and show that the constrained efficient allocations can be decentralized through a competitive asset market where households only

[4]For example, in the limited commitment economy of Krueger and Perri (2006), an increase in the size of idiosyncratic income risk reduces the value of autarky (and incentives to default), and increases equilibrium risk sharing, a force that contains the rise in consumption inequality caused by higher income uncertainty.

trade a non-contingent bond.[5]

The canonical example of the exogenous incomplete market approach is the so-called *bond economy*, an environment where households are only allowed to trade a one period non state-contingent bond. This arrangement is reminiscent of a standard deposit/loan contract in the real world, where a saver receives an interest on their deposits every period, and a borrower repays interests on their loan every period without ever defaulting.[6] To fully understand the ad-hoc restrictions on market structure that we impose to obtain the bond economy, start from complete markets and consider the sequential formulation version of the Arrow-Debreu budget constraint (11.2) holding at every history $\omega^t \in \Omega^t$ that we have studied in Section 7.4:

$$c_{i,t}\left(\omega^t\right) + \sum_{\omega_{t+1}\in\Omega} q_t\left(\omega_{t+1},\omega^t\right) a_{i,t+1}\left(\omega_{t+1},\omega^t\right) = y_{i,t}\left(\omega^t\right) + a_{i,t}\left(\omega^t\right) \tag{11.7}$$

where $a\left(\omega_{t+1},\omega^t\right)$ is an Arrow security purchased at date t and state ω^t that pays one unit of consumption if state ω_{t+1} occurs next period, $q\left(\omega_{t+1},\omega^t\right)$ is the price of such Arrow security, and $a_{i,t}\left(\omega^t\right)$ are all the Arrow securities purchased at $t-1$ which pay in the current realized state ω^t. In the bond economy, we force agents to trade only a non state-contingent asset. The budget constraint (11.7) is therefore replaced by the more restrictive

$$c_{i,t}\left(\omega^t\right) + q_t\left(\omega^t\right) a_{i,t+1}\left(\omega^t\right) = y_{i,t}\left(\omega^t\right) + a_{i,t}\left(\omega^{t-1}\right), \tag{11.8}$$

where $a_{t+1}\left(\omega^t\right)$ is a bond purchased at date t and state ω^t that pays one unit of consumption next period, independently of the realization of the state ω_{t+1}, and $q_t(\omega^t)$ is the price of such bond.

As the terminal condition, for now we only impose the no Ponzi game (nPg) condition and do not impose any further borrowing constraints. The nPg condition for this market structure is

$$\lim_{t\to\infty}\left(\prod_{s=0}^{t} q_s\left(\omega^s(\omega^t)\right)\right) a_{i,t+1}\left(\omega^t\right) \geq 0 \tag{11.9}$$

for all $\omega^t \in \Omega^t$, where $\omega^s(\omega^t)$ represents the sub-history of ω^t up to period s. A similar condition (for the deterministic case) has shown up in Section 5.3.1. Optimality implies the transversality condition (TVC) $\lim_{t\to\infty}\left(\prod_{s=0}^{t} q_s\left(\omega^s(\omega^t)\right)\right) a_{i,t+1}\left(\omega^t\right) \leq 0$ for all $\omega^t \in \Omega^t$, since a household with a positive limiting value of their wealth can improve their welfare by dissaving and consuming a bit extra.[7] Combining these two inequalities, we obtain the condition

$$\lim_{t\to\infty}\left(\prod_{s=0}^{t} q_s\left(\omega^s(\omega^t)\right)\right) a_{i,t+1}\left(\omega^t\right) = 0 \quad \text{for all} \quad \omega^t \in \Omega^t. \tag{11.10}$$

[5] Broer, Kapička, and Klein (2017) obtain similar implications from an economy which combines limited enforcement of contracts and private information about earnings.

[6] In reality such contracts are intermediated by banks. To the extent that the financial sector is competitive and banks solve a static maximization problem, financial intermediaries play no interesting role in the model and can be ignored. Chapter 19 discusses dynamic models of financial frictions where banks play a crucial role in determining equilibrium allocations.

[7] These concepts were introduced in Section 4.3.1.

To simplify the analysis, in what follows we assume away fluctuations in the aggregate endowment $Y_t(\omega^t)$ which implies that the bond price is a constant, $p_t(\omega^t) = p$ and that we can omit the dependence on histories. Since the bond pays one unit of consumption in the next period, its rate of return is $r = 1/p - 1$. We can therefore reformulate the individual budget constraint of the bond economy (11.8) as

$$a_{i,t+1} = (1+r)(y_{i,t} + a_{i,t} - c_{i,t}). \tag{11.11}$$

This budget constraint (which reflects the market structure of the bond economy) is one of the cornerstones of the analysis of consumption and saving behavior in modern macroeconomics.[8] In this economy there are no explicit insurance markets, but, beyond borrowing when allowed, saving and dissaving act as a mechanism for *self-insurance*.[9] Throughout the rest of the chapter, we'll maintain this market structure.

11.3 Income fluctuation problems

We begin with a partial equilibrium analysis of the intertemporal decision problem of an infinitely-lived household who is subject to fluctuations in labor income, and every period must decide how much to consume and how much to save in a risk-free non-state contingent asset, possibly subject to a borrowing limit. The household takes the interest rate as given.

We analyze this *income fluctuation problem* (in the language of Schechtman and Escudero, 1977) first in the deterministic case and then in the stochastic case. A special case of this model, when the utility function is quadratic, is Hall (1978)'s formulation of Friedman (1957)'s permanent income hypothesis which displays certainty equivalence. We then generalize the model to settings where risk matters, either because of the presence of binding liquidity constraints or because the utility function displays prudence. For each of these models, we characterize the marginal propensity to consume.

11.3.1 Deterministic case

We begin by abstracting from income uncertainty. Consider the problem of an individual who faces deterministic (i.e., perfectly known ex-ante) income fluctuations $\{y_t\}_{t=0}^{\infty}$, has to choose optimally how to allocate consumption c_t over time, and can only save through a risk free bond a_t. This individual solves

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{11.12}$$

[8] Note the timing convention implicit in the way we wrote this budget constraint: income is paid and consumption is chosen at the beginning of the period, and thus interests accrue on savings defined as $y_t + a_t - c_t$. The alternative timing convention, which we'll sometimes use in this chapter, is that income is paid and consumption decisions are made at the end of the period, which leads to the formulation of the budget constraint: $a_{t+1} = y_t + (1+r)a_t - c_t$.

[9] The expression "self-insurance" refers to the fact that individuals are insuring against future shocks by dissaving and saving, i.e. by trading intertemporally with themselves, and not by trading state-contingent insurance contracts with others, as such contracts are not available in this market structure.

subject to

$$a_{t+1} = (1+r)(y_t + a_t - c_t).$$

This problem is analogous to the consumption-saving model from Chapter 4, but extended to incorporate time-varying deterministic income levels y_t.

By solving the Lagrangean associated to this problem, it is easy to see that the household first order condition yields

$$\frac{u'(c_t)}{\beta u'(c_{t+1})} = 1 + r. \tag{11.13}$$

This *consumption Euler equation* has the standard interpretation: the marginal rate of substitution between consumption today and consumption next period (intended as two different goods) equals the price of consumption today relative to the price of consumption next period, or the interest rate.[10] It also has a variational interpretation: the value of a unit of consumption today is its marginal utility. Shifting such unit to next period yields $(1+r)$ units valued at the discounted marginal utility tomorrow. Optimality requires the household to be indifferent, hence the equal sign.

Because of the strict concavity of u, the Euler equation implies that the slope of the optimal individual consumption profile between t and $t+1$ is increasing in β and in r. Both higher patience and higher return on saving induce the household to save more today and postpone consumption to the future, which tilts upward the consumption profile.[11] Turning to the special case of CRRA utility with curvature parameter σ, (11.13) becomes

$$\frac{c_{t+1}}{c_t} = [\beta(1+r)]^{\frac{1}{\sigma}}. \tag{11.14}$$

This formulation clarifies that the extent to which variation in β and r translates into steeper or flatter consumption paths depends on the elasticity of intertemporal substitution, $EIS = 1/\sigma$. See Section 4.2.4 for details on the derivation of the EIS.

To sum up, in this environment, households want to smooth consumption with respect to deterministic income fluctuations, and thus they save when income is high relative to its mean and dissave when it is low. The extent of consumption smoothing (i.e., how close c_t and c_{t+1} are to each other) depends on the product $\beta(1+r)$ and on the willingness of households to substitute intertemporally, determined by $1/\sigma$. The closer $\beta(1+r)$ is to 1 and the lower the elasticity of substitution, the more consumption is smoothed across periods.

[10] To see why the interest rate is the intertemporal price of consumption, rewrite the budget constraint (11.8) in nominal terms with the price of the final good multiplying quantities at t and $t+1$

$$p_{t+1}a_{t+1} = p_t a_t + p_t y_t - p_t c_t,$$

then divide through by p_{t+1} and compare with (11.8).

[11] These statements are about the relative consumption across the two periods. Whether the *level* of consumption at t increases or decreases depends on the relative strength of income and substitution effect due to the increase in r (and thus on the level of wealth a_t), as for any other change in relative prices.

Marginal propensities to consume

Let $R \equiv 1 + r$ and iterate forward the budget constraint to obtain, after imposing condition (11.10)

$$c_t + \frac{1}{R} c_{t+1} + \frac{1}{R^2} c_{t+2} + ... = a_t + \sum_{j=0}^{\infty} \left(\frac{1}{R}\right)^j y_{t+j}.$$

Using the Euler equation (11.14) to substitute c_{t+j} for all $j > 0$ on the left hand side as a function of c_t, and collecting terms, we arrive at:

$$c_t = \left(1 - R^{-1} (\beta R)^{\frac{1}{\sigma}}\right) \left[a_t + \sum_{j=0}^{\infty} \left(\frac{1}{R}\right)^j y_{t+j}\right]. \tag{11.15}$$

This expression is useful to introduce the concept of *marginal propensity to consume* (MPC). The MPC out of wealth a_t (or, equivalently, out of a transitory change in income y_t) is defined as $\partial c_t / \partial a_t$.[12] Differentiating (11.15), we obtain that

$$MPC = 1 - R^{-1} (\beta R)^{\frac{1}{\sigma}}. \tag{11.16}$$

Two special cases are of interest. First, recall that in the equilibrium of a representative agent model without growth, $\beta R = 1$ and thus $MPC = 1 - \beta$. Let $\beta \equiv 1/(1 + \rho)$, where ρ is the discount rate. Since ρ is small relative to 1, $MPC \simeq \rho$.[13] Thus, in the representative agent model, the marginal propensity to consume is approximately equal to the discount rate. Second, assuming log-utility ($\sigma = 1$) in the individual problem, without imposing $\beta R = 1$, by following similar steps one obtains $MPC \simeq r$.

Expression (11.15) shows that in this simple model optimal consumption is linear in wealth. In addition, equation (11.15) is an incarnation of Friedman's permanent income hypothesis: the term in the square bracket is the sum of financial wealth a_t plus human wealth (the present value of future income), i.e. total wealth. Optimal consumption equals a constant fraction of total wealth. This equation also illustrates one of the key concepts in Friedman's theory of consumption: the marginal propensity to consume out of transitory and permanent changes in income are different. Consider the case $\beta R = 1$ where the MPC out of transitory income is $1 - R^{-1}$. From (11.15) , it is easy to see that a permanent change in income of one unit, i.e. an increase of one unit in y_{t+j} for all $j \geq 0$, leads to a change in human wealth equal to $1/\left(1 - R^{-1}\right)$ and, thus a change in consumption equal to exactly 1 unit. Thus, the MPC out of permanent income is 1, and much larger than the MPC out of transitory income. Next, we explore the permanent income hypothesis in more detail.

11.3.2 Permanent income hypothesis

We now reintroduce income uncertainty and rewrite the household problem with the conditional expectation operator in the objective function as

$$\max_{\{c_t\}_{t=0}^{\infty}} \mathbb{E}_t \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{11.17}$$

[12] By transitory we mean a change in y_t which leaves unchanged income y_{t+j} at any $j > 0$.

[13] Another way of deriving this result is to define $\beta = \exp(-\rho)$ and use the approximation $\exp(x) \simeq 1 + x$.

subject to

$$a_{t+1} = (1+r)(y_t + a_t - c_t).$$

We also make two additional assumptions: quadratic utility

$$u(c_t) = b_1 c_t - \frac{1}{2} b_2 {c_t}^2, \qquad b_2 > 0, \quad c_t < b_1/b_2,$$

and $\beta R = 1$. From the consumption Euler equation implied by (11.17), jointly with these two assumptions, we obtain

$$c_t = \mathbb{E}_t c_{t+1}. \tag{11.18}$$

This is the well known result of Hall (1978) that consumption is a martingale (or a random walk).[14] From the law of iterated expectations and the martingale property of the optimal consumption allocation:

$$\mathbb{E}_t c_{t+2} = \mathbb{E}_t \left[\mathbb{E}_{t+1} c_{t+2}\right] = \mathbb{E}_t c_{t+1} = c_t$$

and, more in general:

$$\mathbb{E}_t c_{t+j} = c_t, \text{ for any } j \geq 0. \tag{11.19}$$

If we iterate forward J times on budget constraint in (11.17), and apply the conditional expectations to deal with uncertain future realizations of income and consumption, we obtain

$$\sum_{j=0}^{J} \left(\frac{1}{R}\right)^j \mathbb{E}_t c_{t+j} = a_t + \sum_{j=0}^{J} \left(\frac{1}{R}\right)^j \mathbb{E}_t y_{t+j} - \left(\frac{1}{R}\right)^{J+1} \mathbb{E}_t a_{t+J+1}$$

Taking the limit as $J \to \infty$ and using condition (11.10), the last term goes to zero. Using the martingale property (11.19) into the left hand side, we obtain

$$c_t = \frac{r}{1+r}\left[a_t + \sum_{j=0}^{\infty} \left(\frac{1}{1+r}\right)^j \mathbb{E}_t y_{t+j}\right]. \tag{11.20}$$

This expression illustrates the stochastic version of the permanent income hypothesis: optimal consumption is, again, linear and equals the annuity value of total wealth, i.e. financial wealth plus human wealth.[15]

If one assumes that income is deterministic and solves (11.17) , one obtains $c_{t+1} = c_t$ from the Euler equation and, by iterating forward on the budget constraint,

$$c_t = \frac{r}{1+r}\left[a_t + \sum_{j=0}^{\infty} \left(\frac{1}{1+r}\right)^j y_{t+j}\right] \tag{11.21}$$

[14] A stochastic process $\{x_t\}$ is a random walk when it satisfies, at every t, $\mathbb{E}_t x_{t+j} = x_t$ for any $j > 0$.

[15] The annuity value is exactly $r/(1+r)$, i.e. that portion of wealth that, when consumed every period, keeps asset holdings constant. To see this, abstract from income y_t, and note that

$$a_{t+1} = (1+r)(a_t - c_t) = (1+r)\left(a_t - \frac{r}{1+r} a_t\right) = a_t.$$

Comparing (11.21) to (11.20) demonstrates that consumption satisfies *certainty equivalence*: in order to obtain the solution of the stochastic problem (11.17), it suffices solving the deterministic problem and applying conditional expectations to the exogenous variables $\{y_{t+j}\}_{j=0}^{\infty}$ in place of the variables themselves. Put differently, no higher moment of the income process, beyond the mean, matters for the dynamics of consumption. This property descends directly from the linear-quadratic structure of the problem. Any deviation from quadratic objective and linear constraints breaks certainty equivalence.

From (11.20), the change in consumption at time t equals

$$\Delta c_t = c_t - c_{t-1} = c_t - \mathbb{E}_{t-1} c_t = \frac{r}{1+r} \left[\varpi_t - \mathbb{E}_{t-1} \varpi_t \right], \tag{11.22}$$

where ϖ_t is total wealth defined as

$$\varpi_t \equiv a_t + \sum_{j=0}^{\infty} \left(\frac{1}{1+r} \right)^j \mathbb{E}_t y_{t+j},$$

and the last term on the right hand side is the innovation, or unexpected change, in permanent income at time t, i.e. the difference between realization and conditional expectation

$$\begin{aligned} \varpi_t - \mathbb{E}_{t-1} \varpi_t &= a_t - \mathbb{E}_{t-1} a_t + \sum_{j=0}^{\infty} \left(\frac{1}{1+r} \right)^j \left[\mathbb{E}_t y_{t+j} - \mathbb{E}_{t-1} \left(\mathbb{E}_t y_{t+j} \right) \right] \\ &= \sum_{j=0}^{\infty} \left(\frac{1}{1+r} \right)^j \left(\mathbb{E}_t - \mathbb{E}_{t-1} \right) y_{t+j}, \end{aligned} \tag{11.23}$$

where we have used the law of iterated expectations $\mathbb{E}_{t-1} \left(\mathbb{E}_t y_{t+j} \right) = \mathbb{E}_{t-1} y_{t+j}$, and the fact that $a_t = \mathbb{E}_{t-1} a_t$, since there is no uncertainty at time t (after y_t is realized) about the evolution of wealth into $t+1$. Combining (11.22) and (11.23), we arrive at

$$\Delta c_t = \frac{r}{1+r} \sum_{j=0}^{\infty} \left(\frac{1}{1+r} \right)^j \left(\mathbb{E}_t - \mathbb{E}_{t-1} \right) y_{t+j}. \tag{11.24}$$

This equation contains another useful result: under the permanent income hypothesis, the change in consumption between $t-1$ and t is proportional to the revision in expected future income due to the new information (the "news") accruing in that same time interval.

Permanent and transitory income shocks

To make further progress, we need to make some assumptions on the income process. We choose a specification that is very common in labor economics, at least since Abowd and Card (1989). We model labor income as the sum of two orthogonal components, y_t^p which follows a martingale with i.i.d. innovation (or shocks) v_t, and u_t which is also an i.i.d. shock

$$\begin{aligned} y_t &= y_t^p + u_t, \\ y_t^p &= y_{t-1}^p + v_t. \end{aligned} \tag{11.25}$$

In addition, we assume that $\mathbb{E}(v_t) = \mathbb{E}(u_t) = 0$ and that the two shocks are orthogonal, $u_t \perp v_\tau$ for all pairs (t, τ). If we let x_t denote either shock, our assumptions imply that $\mathbb{E}_t(x_{t-j}) = x_{t-j}$ for $j \geq 0$, and $\mathbb{E}_t(x_{t+j}) = 0$ for $j > 0$.

Combining these two equations in (11.25), we obtain the representation

$$y_t = y_{t-1} + u_t - u_{t-1} + v_t. \tag{11.26}$$

Using this income process into (11.24), after some algebra, we obtain

$$\Delta c_t = \frac{r}{1+r} u_t + v_t, \tag{11.27}$$

which establishes that households adjust their consumption responding to the annuity value of transitory shocks and to the full value of permanent shocks. This finding is analogous to the one for the deterministic economy: the pass-through of income shocks to consumption depends on their expected duration. Remarkably, this version of the bond economy is quite close to full insurance with respect to transitory shocks. In conclusion, borrowing and saving through a non-state contingent asset offers ample opportunity for consumption smoothing as long as shocks are not too persistent.

Using the PIH to learn about the nature of the rise in inequality

Income inequality increased substantially in the 1980s and the 1990s in many developed countries. Much of the public, including many policymakers and economists, interpreted this trend as indicating widening differentials in standard of living across households. This interpretation, however, is open to the criticism that current income may not reflect the long-run level of resources available to a household, and hence their welfare.

Blundell and Preston (1998) showed how one can use the permanent income hypothesis, together with data on the joint cross-sectional distribution of income and consumption, to learn whether the rise in cross-sectional income inequality is of a transitory nature and hence not too worrisome, or of a permanent nature and thus detrimental for inequality in household welfare.

Consider the model of Section 11.3.2. From equations (11.26) and (11.27), we have that the evolution of income and optimal consumption for an individual i at time t are given by

$$\begin{aligned} y_{i,t} &= y_{i,t-1} + u_{i,t} - u_{i,t-1} + v_{i,t} \\ c_{i,t} &= c_{i,t-1} + \frac{r}{1+r} u_{i,t} + v_{i,t}. \end{aligned}$$

Now, compute the cross-sectional variance of consumption and the cross-sectional covariance between consumption and income for all individuals belonging to a cohort k, assuming that $r \simeq 0$. Then, one obtains:

$$\Delta var_{k,t}(c) = \Delta covar_{k,t}(c, y) \simeq var_t(v). \tag{11.28}$$

In other words, by tracing the change over time of the within-cohort variance of consumption, or covariance between consumption and income, one can estimate the change over time in the variance of the permanent component of income. Blundell and Preston concluded that the bulk of the rise in UK income inequality was driven by the permanent component, a result that is consistent with the idea that skill-biased technical change (and rising college wage premium), is a key driving force of the recent changes in income distribution. We discuss skill-biased technical change in Section 21.2.2. Although based on a different methodology, these empirical findings are reminiscent of those in Attanasio and Davis (1996) discussed in Section 11.2.2. In particular, they also represent a rejection of full insurance since, from the perspective of the efficient risk sharing hypothesis, consumption should not react to idiosyncratic income shocks, no matter their persistence.

Saving for the rainy days

Define household savings s_t as capital income plus labor income net of consumption expenditures

$$s_t = \frac{r}{1+r} a_t + y_t - c_t. \tag{11.29}$$

Combining (11.29) with (11.20) we obtain an expression for saving only as a function of current and future expected income

$$s_t = y_t - \frac{r}{1+r} \sum_{j=0}^{\infty} \left(\frac{1}{1+r} \right)^j \mathbb{E}_t y_{t+j}.$$

Unfolding this summation on the right hand side, we obtain

$$\begin{aligned} s_t &= y_t - \frac{r}{1+r} y_t - \frac{r}{1+r} \left[\left(\frac{1}{1+r} \right) \mathbb{E}_t y_{t+1} + \left(\frac{1}{1+r} \right)^2 \mathbb{E}_t y_{t+2} + ... \right] \\ &= \frac{1}{1+r} y_t - \frac{r}{1+r} \left(\frac{1}{1+r} \right) \mathbb{E}_t y_{t+1} - \frac{r}{1+r} \left[\left(\frac{1}{1+r} \right)^2 \mathbb{E}_t y_{t+2} + ... \right] \\ &= -\frac{1}{1+r} \mathbb{E}_t \Delta y_{t+1} + \left(\frac{1}{1+r} \right)^2 \mathbb{E}_t y_{t+1} - \frac{r}{1+r} \left[\left(\frac{1}{1+r} \right)^2 \mathbb{E}_t y_{t+2} + ... \right] \end{aligned}$$

and, using the same approach on terms $t+j$ with $j > 1$ we obtain

$$s_t = -\sum_{j=1}^{\infty} \left(\frac{1}{1+r} \right)^j \mathbb{E}_t \Delta y_{t+j}.$$

This expression shows that savings are equal to the discounted sum of expected declines in income. If a household expects income to fall in the future, in anticipation they will save and these savings will allow them to smooth consumption once income actually declines. For example, under a mean reverting income process, households who experience a sequence of positive (above mean) income shocks will save because they understand that this abundance is only transient. Similarly, those who experience a sequence of negative (below mean) income shocks will dissave. Campbell (1987) calls this behavior *saving for the rainy day*. Or, dissaving in the expectation of a sunny day.

11.3.3 Borrowing constraints

To derive the results in the previous section, we have abstracted from borrowing constraints. But, to what extent can we safely ignore borrowing limits, as if they were never binding? The answer depends on the income process.

Suppose we impose the ad-hoc no-borrowing constraint $a_{t+1} \geq 0$ on the household problem. When income y_t is a random walk, from (11.20) we obtain

$$c_t = \frac{r}{1+r} a_t + y_t$$

since $\mathbb{E}_t y_{t+j} = y_t$ for all $j \geq 0$. Substituting this expression for consumption into the budget constraint yields $a_{t+1} = a_t$. In this case, wealth is constant, so if a households starts with a positive wealth level, the zero debt limit will never bind.

Suppose now that y_t follows an i.i.d. shock with mean $\bar{y}$. From (11.20) we have

$$c_t = \frac{\bar{y}}{1+r} + \frac{r}{1+r} (a_t + y_t) \tag{11.30}$$

since $\mathbb{E}_t y_{t+j} = \bar{y}$. Substituting into the budget constraints yields $\Delta a_{t+1} = y_t - \bar{y}$. In this case, wealth follows a random walk without drift, which means that starting from any initial level of wealth there is always a positive probability that at some point in the future the household will hit the borrowing limit in finite time. For example, consider an individual who receives an income realization equal to $y_t < (1-r)\bar{y}$ and whose initial wealth is $a_t < r\bar{y}$. It is easy to see that if the individual wanted to consume its optimal unconstrained level in equation (11.30), they would enter the next period with negative wealth, which would violate the no-borrowing limit. As a result, the individual would have to consume below its unconstrained level. More broadly, the borrowing constraint is likely to bind whenever a_t is small, y_t follows a mean-reverting process, and the individual is hit by a low enough shock (relative to its mean). Because income is expected to revert back to its higher mean, consumption smoothing dictates that the household should dissave and keep their consumption high. In some states, however, dissaving is not enough, borrowing would be necessary and the constraint binds. In this case, the optimal consumption choice is constrained.

These examples highlight the fact that, in general, borrowing limits cannot be ignored and their presence affects optimal consumption and saving choices, as we explain next. We now examine the more general consumption-saving problem with stochastic income y_t and an ad-hoc zero-borrowing constraint.

Consider the problem of a household facing a no-borrowing constraint. Stated in a sequential formulation, we have

$$\max_{\{c_t, a_{t+1}\}_{t=0}^{\infty}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t) \tag{11.31}$$

subject to

$$a_{t+1} = R(a_t + y_t - c_t)$$

and

$$a_{t+1} \geq 0,$$

where $u' > 0$ and $u'' < 0$. The first order condition of this problem yields the modified Euler equation:

$$u'(c_t) = \beta R \mathbb{E}_t [u'(c_{t+1})] + \lambda_t,$$

where λ_t is the multiplier on the borrowing constraint at date t. Because $\lambda_t \geq 0$, the Euler equation can be rewritten as

$$u'(c_t) \geq \beta R \mathbb{E}_t [u'(c_{t+1})], \tag{11.32}$$

where the strict inequality holds when the constraint is binding. Assume $\beta R = 1$. If the constraint is not binding, the household's choice is interior, dictated by the Euler equation which states that marginal utility today is equated to expected marginal utility next period. If, instead, the constraint is binding, current marginal utility is higher and consumption is lower: the household problem has a corner solution and is not determined by the Euler equation, but by the borrowing limit and the budget constraint. Since $a_{t+1} = 0$, from the budget constraint, we obtain $c_t = a_t + y_t$, i.e. the household consumes all their cash in hand. The multiplier λ_t plays a role akin to the interest rate, in that it increases the cost of consuming today relative to postponing consumption to the future (Hall, 1988b).

There are two key differences between the model without and with borrowing constraints. The first one is that, absent borrowing constraints, what matters for current consumption is only the expected discounted value of labor income, not its timing (see equation (11.20)). In the presence of borrowing constraints, instead, the timing of income matters for consumption. Consider two random income streams with the same discounted expected value, one that is decreasing and another one that is increasing: a household facing the latter is more likely to be constrained and their consumption path may differ from the consumption path of a household facing the decreasing stream. The second difference is that, for a constrained household, the marginal propensity to consume is 1 for both transitory and permanent shocks: no matter what the duration of the shock is, a small change in current income increases current consumption one for one. Thus constrained (or hand-to-mouth) households display higher MPC out of transitory income relative to unconstrained ones.

Natural borrowing limit

So far, when modelling borrowing constraints we have specified ad-hoc limits of the type $a_{t+1} \geq -\underline{a}$. Another type of constraint is the so called "natural" borrowing limit. In Section 4.3.1, we discussed the notion of the natural borrowing limit in the context of a deterministic model. In the Online Appendix to Chapter 4, we have shown that, in the deterministic model, the natural borrowing limit is equivalent to the no-Ponzi-game (nPg) condition. In what follows, we extend this notion to an environment with uncertainty. Here, the natural borrowing limit is the lowest asset position consistent with non-negative consumption in every state and every period.

Start from the case where $\{y_t\}_{t=0}^{\infty}$ is deterministic. Iterating forward on the budget constraint and imposing $c_t \geq 0$ for all t, we obtain

$$a_{t+1} \geq -\sum_{j=0}^{\infty} \left(\frac{1}{1+r}\right)^j y_{t+1+j}.$$

Assume now that income is stochastic and the lowest possible realization is $y_{\min}$, always occurring with positive probability. Then, to guarantee positive consumption at every t, debt cannot exceed the present value of $y_{\min}$, and thus the natural borrowing limit becomes

$$a_{t+1} \geq -\left(\frac{1+r}{r}\right) y_{\min}.$$

Borrowing up to this value, or more, would mean that there is a positive probability that a state occurs where consumption must be zero. This statement is easy to prove. Suppose the household has borrowed exactly up to this limit, i.e. $a_t = -\left(\frac{1+r}{r}\right) y_{\min}$ and they are hit by an income realization $y_t = y_{\min}$. Then from the budget constraint:

$$c_t = \underbrace{-\left(\frac{1+r}{r}\right) y_{\min}}_{a_t} + y_{\min} - \frac{a_{t+1}}{1+r} \leq -\frac{1}{r} y_{\min} - \frac{1}{1+r} \underbrace{\left[-\left(\frac{1+r}{r}\right) y_{\min}\right]}_{\text{max that can be borrowed}} = 0$$

It follows that, if the utility function satisfies the Inada condition $u(0) = -\infty$, an optimizing consumer will never borrow up to the natural borrowing limit because doing that is inconsistent with expected utility maximization. Thus, one can always safely assume an interior solution for the Euler equation. This is, instead, not true for ad-hoc debt limits which always affect the optimal solution in some region of the state space, as explained above. Finally, note that if $y_{\min} = 0$, then the no-borrowing constraint is also the natural borrowing limit.[16]

11.3.4 Precautionary saving

The term *precautionary saving* refers the additional amount of saving a household chooses in response to a rise in the uncertainty about future income. The certainty equivalence property of consumption allocation under quadratic utility implies that the consumption function is linear in wealth and in the expected present value of future income. Thus, a mean preserving spread of the income shock distribution (i.e., an increase in dispersion around the same mean) does not impact saving: no precautionary motive for saving is present in this environment.

In what follows we establish two forces that lead to precautionary saving behavior. The first one is the presence of occasionally binding borrowing constraints, and the second is prudence. Prudence is a property of the utility function that mathematically corresponds to a positive third derivative, i.e. $u''' > 0$. Under both forces a rise in income uncertainty leads to a rise in the level of savings. As first uncovered by Zeldes (1989) through numerical simulations, in these settings the consumption policy function is concave in wealth, a property that has important implications for the marginal propensity to consume.

[16] Throughout this discussion, we have assumed that $r > 0$. As we will see in Section 11.4, $r < 0$ can be an equilibrium outcome of economies populated by a continuum of households facing an income fluctuation problem. In such a case, it is convenient (and realistic) to assume the existence of a wedge $\chi > 0$ between the lending rate r and the borrowing rate r^b which can be interpreted as a linear intermediation cost of the financial sector. The interest rate entering the natural borrowing limit is $r^b = r + \chi$, and if χ is large enough, $r^b > 0$ and the natural borrowing limit is well defined again.

Precautionary motive with borrowing constraints

To isolate the role of borrowing constraints for precautionary saving, consider again the quadratic specification for period utility, which does not display prudence because $u''' = 0$. Continue assuming that and $\beta R = 1$. Suppose households face a borrowing limit $a_{t+1} \geq -\underline{a}$. Then,

$$c_t = \begin{cases} \mathbb{E}_t c_{t+1} & \text{if } a_{t+1} > -\underline{a} \\ y_t + a_t + \frac{\underline{a}}{R} & \text{if } a_{t+1} = -\underline{a} \end{cases}$$

The first line describes the optimal unconstrained intertemporal consumption allocation when the constraint is not binding, and the second line the consumption allocation if the constraint is binding, in which case the individual consumes all their resources. The above pair of conditions can be written in compound form as

$$c_t = \min \left\{ y_t + a_t + \frac{\underline{a}}{R}, \mathbb{E}_t c_{t+1} \right\}. \tag{11.33}$$

Assume the constraint is not already binding at date t. Then

$$c_t = \mathbb{E}_t c_{t+1} = \mathbb{E}_t \left[\min \left\{ y_{t+1} + a_{t+1} + \frac{\underline{a}}{R}, \mathbb{E}_{t+1} c_{t+2} \right\} \right]. \tag{11.34}$$

Suppose that the uncertainty about y_{t+1} increases but its mean does not change. Very low realizations of income y_{t+1} become more likely, and thus the borrowing constraint is also more likely to bind next period. This reduces the value of $\mathbb{E}_t \left[\min \left\{ y_{t+1} + a_{t+1} + \frac{\underline{a}}{R}, \mathbb{E}_{t+1} c_{t+2} \right\} \right]$ and of current consumption c_t. As a result, savings increase. Intuitively, when households face borrowing limits which can potentially bind in the future, they understand that if they were to receive bad income realizations, they would be pushed towards the constraint and be forced to consume their income without the ability of smoothing consumption, which reduces their welfare. To prevent this scenario, they save more. This is the first source of precautionary saving motive.

An implication of the presence of borrowing constraints is that the consumption function is concave in wealth, as illustrated in Figure 11.2.

If we fix labor income y_t to a sufficiently low value, it is clear from (11.33) that there exists a level of wealth $a_t = a^*$ below which the constraint binds and $c_t = a_t + y_t + \frac{\underline{a}}{R}$. In this region, the slope of the consumption function with respect to wealth is 1. For sufficiently high levels of a_t, it becomes extremely unlikely (or impossible, depending on the process for y_t) that the constraint will bind in the future and consumption asymptotes to the linear unconstrained solution (11.20) with slope $r/(1+r) < 1$. For intermediate levels of wealth, the borrowing limit could be binding again, and consumption is depressed by the precautionary motive relative to the unconstrained optimum. In this range, the consumption function is strictly concave and the marginal propensity to consume is bracketed between $r/(1+r)$ and 1.

The figure also illustrates what happens when the credit limit tightens from $\underline{a}_1$ to $\underline{a}_2 < \underline{a}_1$. Obviously, households are constrained for a wider range of a_t. For unconstrained households, consumption falls across the board because of the stronger precautionary saving motive, and their marginal propensity to consume (the slope of the consumption function) increases. In the limit as wealth keeps growing, the two consumption functions converge because the borrowing constraint becomes irrelevant.

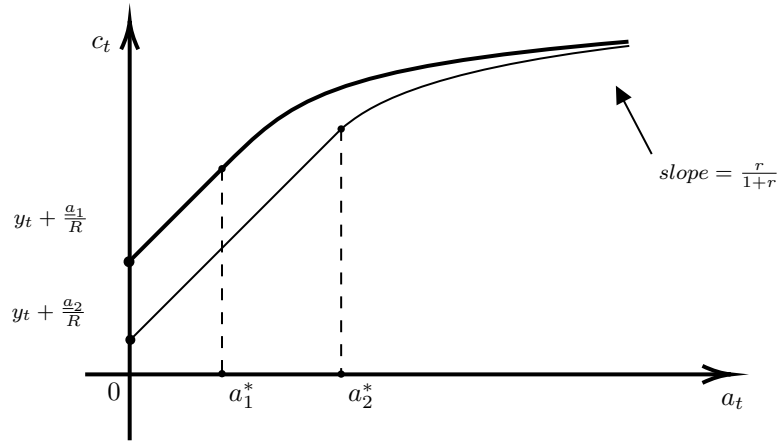


Figure 11.2: Decision rule for consumption in the presence of a borrowing constraint $\underline{a}$ as a function of wealth a_t, for a given realization of income y_t.

Notes: The function is linear, with slope equal to 1, until a^*, after which it becomes concave. For large enough wealth, its slope converges to $\frac{r}{1+r}$. The two lines correspond to different values of the borrowing limit.

An important implication of this discussion is that there is a range of wealth levels for which the individual MPC is larger than in the PIH, either because the constraint binds (and the consumption function has slope equal to 1) or because it might bind with some probability in the future (and the consumption function is strictly concave). We return to this observation in Section 11.4.2.

Precautionary motive with prudence

To illustrate the second source of precautionary saving, consider a simple two-period consumption-saving problem (Leland, 1968; Sandmo, 1970)

$$\max_{\{c_0,c_1,a_1\}} u(c_0) + \beta \mathbb{E}_0 [u(c_1)]$$

subject to

$$c_0 + a_1 = y_0$$

and

$$c_1 = Ra_1 + y_1,$$

where y_0 is given, and y_1 is stochastic. Again, to isolate this second force from the first one we just described, we assume away any borrowing constraint at $t = 0$. The Euler equation for this problem is

$$u'(y_0 - a_1) = \beta R \mathbb{E}_0 [u'(Ra_1 + y_1)] \tag{11.35}$$

an equation in one unknown, a_1. The left-hand-side is increasing in a_1 since $u'' < 0$, and the right-hand-side is decreasing for the same reason, hence the solution for a_1 is uniquely determined. What happens to optimal consumption at $t = 0$ as future income y_1 becomes more risky? Consider a mean-preserving spread of y_1. If u' is convex then, by Jensen's

inequality, the value of the right-hand-side will increase. Graphically (see Figure 11.3), the left-hand-side is unchanged and the right-hand-side shifts upward, inducing a rise in optimal savings a_1.[17] One way to understand this result is that the hike in future consumption uncertainty reduces welfare. Increasing savings today raises the expected value of future consumption which compensates for higher variance. Sibley (1975) and Miller (1974) extend this proof to a multi-period model with a finite T, and i.i.d. income shocks.

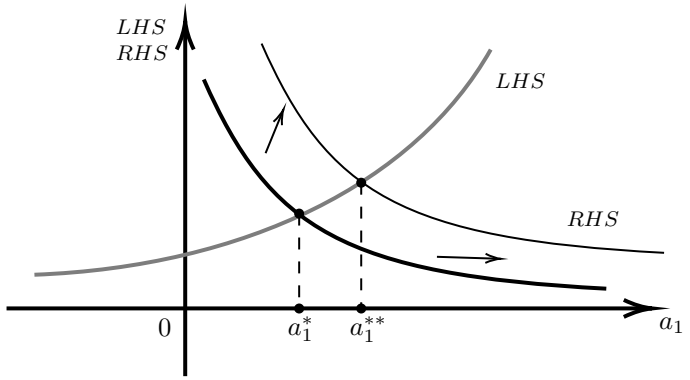


Figure 11.3: Left hand side (LHS) and right hand side (RHS) of the Euler equation (11.35).

Notes:It shows that when a mean-preserving spread in future income occurs, the RHS shifts outward and the optimal amount of saving in period zero a_1 increases.

The convexity of the marginal utility corresponds precisely to the condition $u''' > 0$, or prudence. Prudence is a property of preferences, like risk aversion: risk-aversion refers to the curvature of the utility function. Prudence refers to the curvature of the marginal utility function. Kimball (1990) defines the index of absolute prudence as $-u'''(c)/u''(c)$ and relative prudence as $-u'''(c)\,c/u''(c)$ in a conceptually similar way to the Arrow-Pratt index of absolute and relative risk-aversion. There exists also an interesting relationship between the two: any utility function in the decreasing absolute risk aversion (DARA) class, which includes CRRA, displays prudence. To see this, let $\alpha(c)$ be the coefficient of absolute risk aversion, then

$$\alpha(c) = \frac{-u''(c)}{u'(c)} \Rightarrow \alpha'(c) = \frac{-u'''(c)\,u'(c) + [u''(c)]^2}{[u'(c)]^2}.$$

Since with DARA $\alpha'(c) < 0$, then we have that

$$-u'''(c)\,u'(c) + [u''(c)]^2 < 0 \Rightarrow u'''(c) > \frac{[u''(c)]^2}{u'(c)} > 0.$$

Intuitively, a rise in uncertainty reduces the certainty-equivalent income next period and, with DARA utility, it effectively increases the degree of risk-aversion of the agent, inducing them to save more.

[17] This result is a simple application of Stiglitz and Rothschild (1970)

Role of prudence: an analytical expression

Blanchard and Mankiw (1988) derive an analytical expression that illustrates how risk affects optimal consumption and saving. Take a second-order approximation of $u'(c_{t+1})$ around the value $c_{t+1} = c_t$:

$$u'(c_{t+1}) \simeq u'(c_t) + u''(c_t)(c_{t+1} - c_t) + \frac{1}{2}u'''(c_t)(c_{t+1} - c_t)^2.$$

Substitute this approximation into the right-hand side of the Euler equation $u'(c_t) = \beta R \mathbb{E}_t u'(c_{t+1})$, divide both sides by c_t^2, and use the approximation

$$(\beta R)^{-1} = \frac{1+\rho}{1+r} = \exp\left(\log\frac{1+\rho}{1+r}\right) \simeq 1 + \rho - r.$$

Rearranging terms, we obtain

$$\mathbb{E}_t\left(\frac{c_{t+1} - c_t}{c_t}\right) \simeq EIS(c_t) \cdot (r - \rho) + \frac{1}{2}P(c_t) \cdot \mathbb{E}_t\left[\left(\frac{c_{t+1} - c_t}{c_t}\right)^2\right] \tag{11.36}$$

where

$$EIS(c_t) \equiv -\frac{u'(c_t)}{c_t u''(c_t)}, \text{ and } P(c_t) = -\frac{u'''(c_t)\, c_t}{u''(c_t)}$$

are, respectively, the elasticity of intertemporal substitution and the coefficient of relative prudence. Equation (11.36) features the two determinants of expected consumption growth. The first term captures the standard intertemporal consumption smoothing motive, active as long as u is strictly concave. The second term contains the role of risk, captured by the expected variability of consumption growth (measured by the second uncentered moment): the stronger is relative prudence, the more consumption growth will respond to changes in risk. Higher uncertainty in future consumption growth, for given expected income, implies higher saving through the precautionary motive.

Similarly to the case where debt limits are binding, one can prove that even in absence of constraints, the consumption function is concave in wealth if the utility function belongs to the hyperbolic absolute risk aversion (HARA) class and displays positive third derivative.[18] First, recall that the envelope condition of the household problem yields

$$u'(c(a)) = V'(a) \tag{11.37}$$

where V denotes the value function and where, to ease notation, we have omitted the dependence on labor income. It can be proved that, under general conditions, the value functions inherits the assumptions on u, i.e. it is increasing, concave and differentiable—see the discussions in Section 4.4. Differentiating both sides with respect to a gives

[18] A utility function belongs to the HARA class if its coefficient of absolute risk aversion is an hyperbola (or, equivalently, absolute risk tolerance is an affine function of wealth). Many common utility functions, such as quadratic, CRRA and CARA are special cases of HARA.

$$c'(a) = \frac{V''(a)}{u''(c(a))} > 0, \tag{11.38}$$

which implies that optimal consumption is strictly increasing in wealth. Differentiating one more time, we obtain

$$c''(a) = \frac{V'''(a)\,u''(c(a)) - V''(a)\,u'''(c(a))\,c'(a)}{[u''(c(a))]^2}. \tag{11.39}$$

Using (11.37) and (11.38) into (11.39), we conclude that the consumption function is concave ($c'' \leq 0$) whenever

$$\frac{V'''(a)\,V'(a)}{V''(a)^2} \geq \frac{u'''(c(a))\,u'(c(a))}{u''(c(a))^2}. \tag{11.40}$$

This is the content of Lemma 2 in Carroll and Kimball (1996). If u belongs to the HARA class, then one can show that the right hand side of (11.40) is equal to a constant $\kappa \geq 0$, i.e. prudence is κ times higher than risk aversion.[19] Assuming HARA, Carroll and Kimball (1996) and Jensen (2018) also prove that, under a finite optimization horizon, the inequality in (11.40) is true and therefore the consumption function is concave. Strict concavity arises for $\kappa > 0$, which holds for the CRRA class, with the exception of $\kappa = 1$ corresponding to CARA utility.[20]

Once again, the intuition is that the precautionary saving motive is declining in wealth. Consider the deviation in optimal consumption in the presence of uninsurable risk relative to the no-uncertainty case in which the consumption function is linear. For large enough wealth, precautionary saving approaches zero and the consumption function asymptotes the no uncertainty case. As wealth falls, precautionary saving keeps rising and consumption keeps falling, hence the concavity.

Capital income uncertainty

What if the income uncertainty refers to capital income, i.e. to the rate of return on saving r, instead of labor income? Does an increase in uncertainty still induce more saving? Consider a problem analogous to the one analyzed earlier in this section. The household at date $t = 0$ receives an income y_0 which they can either consume or save, i.e $y_0 = c_0 + a_1$. In the second period, consumption is simply $c_1 = (1 + r)\,a_1$, but $r > -1$ is now stochastic. The Euler equation corresponding to this problem is

$$u'(y_0 - a_1) = \beta \mathbb{E}_0 \left[(1 + r)\,u'((1 + r)\,a_1)\right].$$

Note that the random interest rate is now inside the expectation. As before the left-hand

[19] Specifically, quadratic utility is the knife-edge case, corresponding to $\kappa = 0$, constant absolute risk aversion (CARA) corresponds to $\kappa = 1$, and constant relative risk aversion (CRRA) utility functions satisfy $\kappa > 1$.

[20] In fact, Cantor (1985) proved that the optimal solution to the consumption/saving problem of a CARA consumer facing uninsurable labor income risk (and no borrowing constraints) displays consumption which is linear in wealth. Compared to the no-uncertainty case, uninsurable risk only affects the intercept of the consumption function by lowering it. The slope (i.e., the constant MPC) is the same.

side is increasing in a_1 and the right-hand side decreasing in a_1, hence the solution is unique. What happens to optimal saving a_1 after a mean-preserving spread in r? The answer is not obvious ex-ante. On the one hand, the precautionary saving force would suggest that savings optimally expand. On the other hand, with higher savings, the household becomes even more exposed to risk, and thus reducing saving curtails risk. This latter force was not present when we analyzed labor income risk.

Under prudence, however, the right-hand side is still a convex function of r (it is the product of a linear and a convex function), and thus with this additional assumption we can conclude that precautionary saving behavior emerges even in the face of higher volatility of returns.

Buffer-stock saving

Consider a version of the intertemporal consumption problem where households have a finite horizon, utility is CRRA with curvature parameter $\sigma > 0$, log income follows a stochastic process which is the sum of a permanent shock y_t^P and a transitory shock u_t (as in Section 11.3.2, but in logs), both Normally distributed. In addition, there is positive probability that income will be zero at any t (capturing, e.g., an unemployment shock) and, as a result, the natural borrowing limit is zero. Finally, the discount rate ρ exceeds the interest rate r.[21]

Optimal consumption in this setting, called "buffer-stock model", is fully characterized in a series of papers by Carroll (1992) and carroll2001theory. What makes this model particularly simple to analyze is that the individual state space can be reduced to one state variable only, x_t, the ratio of cash in hand $Ra_t + y_t$ to permanent income y_t^P. Similarly to our derivation of equation (11.36), and focusing on isoelastic utility with curvature parameter σ, one can show that

$$\mathbb{E}_t \Delta \log c_{t+1} \simeq \frac{1}{\sigma}(r - \rho) + \frac{\sigma + 1}{2} var_t (\Delta \log c_{t+1}), \tag{11.41}$$

where var_t denotes the conditional variance, and where $\Delta \log c_t$ replaces the growth rate which appeared in equation (11.36).[22] As $x_t \to 0$, expected consumption growth approaches infinity because c_t approaches zero: when $x_t = 0$ from the budget constraint $-c_t = a_{t+1}/R \geq 0$, and thus $c_t = 0$. As $x_t \to \infty$, uncertainty about future labor income becomes essentially irrelevant and expected consumption growth approaches $\sigma^{-1}(r - \rho) < 0$.

Figure 11.4 illustrates the relation between consumption growth and normalized wealth x. The figure highlights that there exists a level of wealth x^* that is dynamically stable. If $x_t > x^*$ the negative intertemporal saving motive outweighs the precautionary saving motive. Consumption growth is negative, and wealth converges back toward x^*. If $x_t < x^*$, the precautionary motive dominates, consumption growth is positive, and wealth rises toward x^*. In sum, households have a *target level of assets* x^* which is above the level of the no uncertainty case, i.e., they hold additional wealth as a buffer stock. Labor income shocks take households periodically above or below this target level, and households adjust their optimal consumption in order to quickly move back to x^*.

[21] As we prove in Section 11.4, $r < \rho$ is an equilibrium outcome in an economy populated by a continuum of this type of consumers.

[22] For a full derivation, see for example Jappelli and Pistaferri (2017), chapter 6.

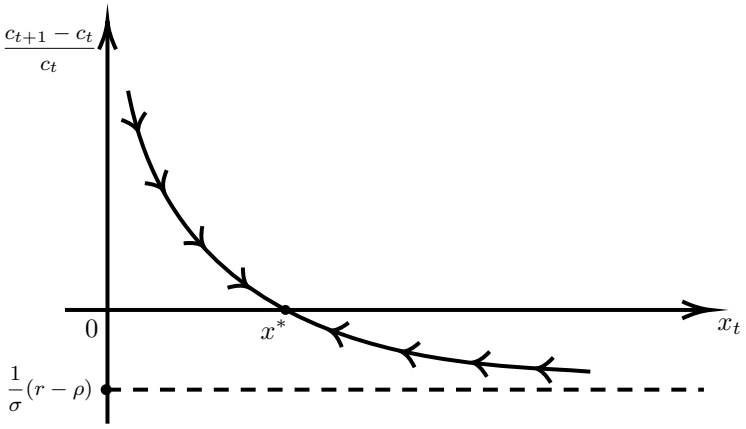


Figure 11.4: Consumption dynamics in the buffer stock model as a function of cash-in-hand x.

Notes:There exists an optimal target level of cash in hand x toward which the household wants to move.

Excess sensitivity and liquidity constraints

According to the permanent income hypothesis, consumption growth between $t-1$ and t should only depend on the innovation in income accruing over the same period. Any past income change should not affect current consumption growth. Empirical tests of this prediction routinely lead to a rejection (e.g., Flavin, 1981) of this hypothesis and to evidence of so-called "excess sensitivity." Deaton (1992a) and Jappelli and Pistaferri (2017)) discuss various econometric difficulties with this type of tests. Campbell and Mankiw (1989) also found evidence of excess sensitivity on aggregate time series data, and interpreted the size of the coefficient on past income as the share of "hand-to-mouth" consumers, i.e. households that are either myopic or subject to liquidity constraints.

Zeldes (1989) links explicitly excess sensitivity tests to the presence of credit constraints. If we generalize equation (11.41) to the case where household face a borrowing limit, we obtain the following Euler equation for consumption:

$$\Delta \log c_{t+1} = \frac{1}{\sigma}(r-\delta) + \frac{\sigma+1}{2}\text{var}_t(\Delta \log c_{t+1}) + \frac{1}{\sigma}\log(1+\lambda_t) + \varepsilon_{t+1}$$

where λ_t is the multiplier on the borrowing limit, and ε_{t+1} is a forecasting error orthogonal to consumption growth. The coefficient on λ_t is positive because when the marginal value of wealth is high, households save more (and consume less) at t, which pushes up consumption growth between t and $t+1$. Note that the multiplier is inherently unobservable, so it acts as an omitted variable.

Zeldes splits his empirical sample of PSID households into low-wealth (likely constrained) and high-wealth (likely unconstrained) households, and adds lagged income as a regressor to that equation to proxy for the multiplier. He argues that, if liquidity constraints bind, one would estimate a significant negative coefficient on lagged income for the low-wealth sample. The reason is that a high income realization relaxes the constraint and lowers the value of the multiplier λ_t. Zeldes' empirical estimates are con-

sistent with the view that liquidity constraints bind for a subgroup of the population, but not for all.

Comparative statics on the consumption function

Figure 11.5 plots the optimal consumption function obtained by solving numerically the income fluctuation problem (11.31). In particular, we have assumed that u is CRRA with risk aversion equal to 2, $\beta = 0.95$, $R = 1.02$, the borrowing limit is zero, and income follows a two-state Markov chain with values $\{0.5, 1.5\}$, and probability that each income state persists into next period equal to 0.8.

The six panels illustrate how the solution to the household problem changes as we change one parameter at the time. Panel (A) shows that households with a low income realization consume less on average for two reasons. First, mechanically, they have less resources. Second, when they are not constrained they are more concerned about hitting the constraint in the future and thus have a stronger precautionary motive. Note that, as wealth grows the households becomes a permanent-income consumer, and the consumption gap converges to the expected difference in permanent labor income. In particular, for wealth large enough the slope of the consumption function is the same. Turning to panel (B), more impatient households (low discounting) consume more and, as long as they are unconstrained, have a higher MPC, even for high values of wealth. Recall the expression for the MPC under full insurance in equation (11.15). Panel (C) shows that, since $\beta R < 1$ households who are more willing to substitute intertemporally consume more. In addition, as seen from (11.15), their MPC is larger even for high levels of wealth. Panel (D) illustrates that households who face a higher rate of return on saving optimally accumulate more assets and consume less for a given level of wealth (but they will be richer on average). Panel (E) shows that a mean-preserving spread in income risk lowers consumption because it expands precautionary saving. For the same reason, the credit constraint binds for a smaller wealth range. Finally, panel (F) depicts optimal consumption under two values for the credit limit. Households who face a loose constraint save less for precautionary reasons. This result indicates the existence of substitutability between precautionary saving and ability to borrow to smooth consumption when needed.

11.3.5 Bounds on wealth accumulation

As discussed, there are two forces that determine saving rates in the stochastic income fluctuation problem: intertemporal and precautionary motive. The latter is always positive. Whereas the first is positive or negative depends on whether βR is above or below one. It is therefore intuitive that unless βR is small enough, households will tend to keep accumulating savings, and an upper bound on wealth might not exist. This result represents a threat to existence of equilibrium once we combine a continuum of households and let them trade in financial markets: if assets can grow without limit, the state space is not compact and an equilibrium might not exist.[23]

It turns out that a sufficient condition for wealth accumulation to be bounded in the stochastic version of the income fluctuation problem is $\beta R < 1$ (Schechtman and Escudero,

[23] A lower bound for the asset space is always guaranteed by the ad-hoc or natural borrowing limit.

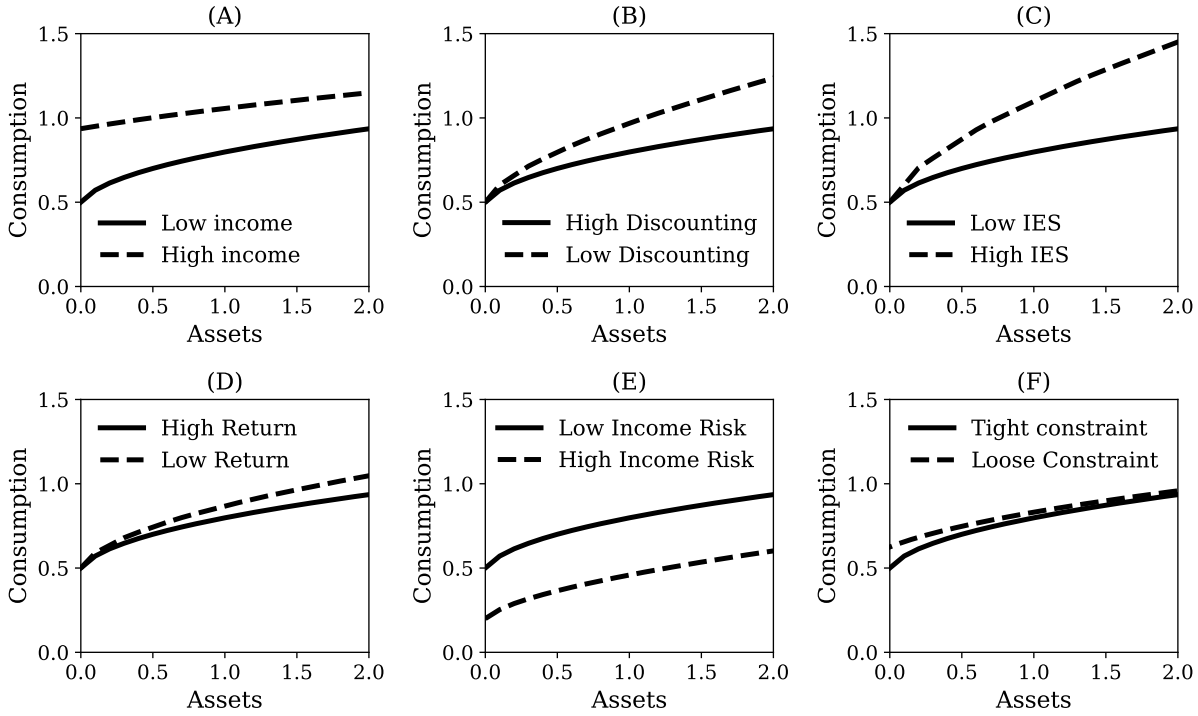


Figure 11.5: Comparative statics of the optimal consumption function with respect to various parameters in the income fluctuations problem.

1977; Clarida, 1987; Huggett, 1997).[24] We now provide an informal version of the proof of this result based on Schechtman and Escudero (1977): besides $\beta R < 1$, the proof assumes i.i.d. income shocks. We use a recursive formulation where 'prime' denotes next-period variables. To avoid confusion, here we use the special notation where $u_c(\cdot)$ is the first derivative of the utility function (the marginal utility of consumption) and u_{cc} is the second derivative of the utility function. Let $x = Ra + y$ be cash in hand, and $c(x)$ the consumption decision rule—a function of the individual state variable x. $c_x(\cdot)$ represents the first derivative of this function. From the Euler Equation:

$$u_c\left(c\left(x\right)\right) = \beta R \mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right], \tag{11.42}$$

where $x' = Ra'\left(x\right) + y'$ is cash in hand next period, given that today's cash in hand is x and that next period income realization is y'. Let $\bar{x}' = Ra'\left(x\right) + \bar{y}$ be the cash in hand associated to the maximum realization of income $\bar{y}$ next period (hence maximum cash in hand next period), given today's cash in hand x. We can write:

$$u_c\left(c\left(x\right)\right) = \beta R \mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right] = \beta R \frac{\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right]}{u_c\left(c\left(\bar{x}'\right)\right)} u_c\left(c\left(\bar{x}'\right)\right). \tag{11.43}$$

Suppose we can show that

$$\lim_{x \to \infty} \frac{\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right]}{u_c\left(c\left(\bar{x}'\right)\right)} = 1. \tag{11.44}$$

[24] The counterpart sufficient condition in the deterministic version of the income fluctuation problem is $\beta R \leq 1$. It is a weaker condition because the precautionary saving motive is absent. See Ljungqvist and Sargent (2018) for a thorough discussion of both the deterministic and the stochastic case.

Then, since $\beta R < 1$ by assumption, for x large enough the Euler equation (11.43) yields

$$u_c\left(c\left(x\right)\right) = \beta R u_c\left(c\left(\bar{x}'\right)\right) < u_c\left(c\left(\bar{x}'\right)\right).$$

Concavity of u and monotonicity of c with respect to x, proved in (11.38) imply that $\bar{x}'\left(x\right) < x$. This would conclude the proof, because we would have demonstrated that cash in hand does not increase forever: when x is large enough, $x' < x$ for sure.

We need to establish conditions under which the limit in (11.44) holds. Consider the marginal utility function $u_c\left(c\left(x'\right)\right)$ and compute a first-order Taylor approximation around $x' = \bar{x}'$:

$$u_c\left(c\left(x'\right)\right) \simeq u_c\left(c\left(\bar{x}'\right)\right) + u_{cc}\left(c\left(\bar{x}'\right)\right) c_x\left(\bar{x}'\right)\left(x' - \bar{x}'\right).$$

Taking expectations of both sides

$$\begin{aligned} \mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right] &\simeq u_c\left(c\left(\bar{x}'\right)\right) - u_{cc}\left(c\left(\bar{x}'\right)\right)\mathbb{E}\left[\bar{x}' - x'\right] c_x\left(\bar{x}'\right) \\ &= u_c\left(c\left(\bar{x}'\right)\right) - u_{cc}\left(c\left(\bar{x}'\right)\right)\mathbb{E}\left[\bar{y} - y'\right] c_x\left(\bar{x}'\right). \end{aligned} \tag{11.45}$$

In the first line we used that $\bar{x}'$ is deterministic since it is implied by the specific income realization $\bar{y}$. In the second line we used the fact that $x' \equiv Ra' + y'$ which implies $\bar{x}' - x' \equiv \bar{y} - y'$. Dividing equation (11.45) by $u_c\left(c\left(\bar{x}'\right)\right)$ we obtain:

$$\frac{\mathbb{E}\left[u_c\left(c\left(x'\right)\right)\right]}{u_c\left(c\left(\bar{x}'\right)\right)} \simeq 1 + \alpha\left(c\left(\bar{x}'\right)\right)\left[\bar{y} - \mathbb{E}\left(y'\right)\right] c_x\left(\bar{x}'\right),$$

where $\alpha\left(c\left(\bar{x}'\right)\right)$ is the coefficient of absolute risk aversion evaluated at consumption level $c\left(\bar{x}'\right)$. Since both $\bar{y} - \mathbb{E}\left(y'\right)$ and $c_x\left(\bar{x}'\right)$ are positive and finite (recall that the c function is concave), a sufficient condition for the limit in (11.44) to hold is that

$$\lim_{x \to \infty} \alpha\left(c\left(x\right)\right) = 0. \tag{11.46}$$

Condition (11.46) requires that absolute risk aversion decreases with wealth, i.e. that u belongs to the DARA class. DARA means that the household is less and less concerned about income uncertainty as they get rich because they are less risk averse, so they will consume more and accumulate less precautionary wealth. As wealth increases, the intertemporal dissaving motive (present because of $\beta R < 1$) eventually overcomes the precautionary saving motive, and wealth will decrease.[25]

To conclude, under DARA the condition $\beta R < 1$ is sufficient to guarantee a bounded asset space in the household problem. In the next section, we'll prove that this inequality is also a feature of the stationary competitive equilibrium in an economy with a continuum of households.

11.4 Heterogeneous-agent incomplete-market models

Now that we have analyzed in some depth the income fluctuation problem, we want to characterize the equilibrium of economies populated by a continuum of households subject to

[25] Huggett (1993a) generalizes this proof to a 2-state Markov chain for the income process, under CRRA utility.

uninsurable income shocks who can only trade a risk-free bond. This class of heterogeneous-agent incomplete-market models has become a workhorse of modern macroeconomics. These models feature an endogenous distribution of consumption and wealth across households. In sharp contrast with its complete-market counterpart, this framework features partial consumption insurance, and mobility within the consumption distribution. In addition, its equilibrium allocations are not socially efficient, which makes policy analysis interesting.

A variety of important questions have been addressed in this framework, for example: what is the fraction of wealth accumulated because of the precautionary motive (Aiyagari, 1994)? How much of the observed wealth inequality can one explain through income inequality (Castaneda, Diaz-Gimenez, and Rios-Rull, 2003)? What are the optimal levels of taxation and public debt (Aiyagari and McGrattan, 1998; Domeij and Floden, 2006; Heathcote, 2005)? How does heterogeneity change the transmission of aggregate shocks compared to the representative agent model (Krusell and Smith, 1998; Kaplan and Violante, 2018)? Can market incompleteness help in generating a large equity premium (Storesletten, Telmer, and Yaron, 2001, Krueger and Lustig, 2010)? Are welfare costs of business cycles higher with incomplete markets (Krusell, Mukoyama, Şahin, and Smith Jr, 2009)?[26]

In what follows, we describe two versions of the model, an endowment economy and a production economy, and discuss existence and uniqueness of the stationary equilibrium.

11.4.1 An endowment economy

Consider the endowment economy introduced in Section 7.6. Time is discrete and indexed by t. There is no aggregate uncertainty. The economy is populated with a continuum of measure one of infinitely lived, ex-ante identical households. Households have time-separable preferences over streams of consumption (the final numeraire good)

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t),$$

where the period utility function u satisfies $u' > 0, u'' < 0$, and the discount factor $\beta \in (0,1)$. The expectation is over future sequences of idiosyncratic shocks, conditional on the information set at time $t = 0$. Each household faces stochastic fluctuations in their endowment of the final good $y_t \in Y$, a finite set. This stochastic process follows a first-order ergodic N-state Markov chain with transition probabilities $\pi(y', y) = \Pr(y_{t+1} = y' \mid y_t = y)$. Since shocks are i.i.d across consumers, a law of large numbers holds: $\pi(y', y)$ is also the fraction of agents in the population subject to this particular transition from y to y' between time t and $t+1$ (Uhlig, 1996). Let $\Pi^*(y)$ be the a unique invariant (i.e. limiting) distribution of endowments. The household budget constraint at time t is

$$c_t + a_{t+1} = Ra_t + y_t.$$

Wealth a_t takes the form of a one-period non-state-contingent asset, with return $R \equiv 1 + r$ independently of the individual realization y_t. At every t, agents face the ad-hoc borrowing

[26] For surveys of applications of this framework, see Heathcote, Storesletten, and Violante (2009); Quadrini and Ríos-Rull (2015); De Nardi and Fella (2017); Benhabib, Bisin, and Luo (2019); Kaplan and Violante (2022).

limit

$$a_{t+1} \geq -\underline{a}$$

where $\underline{a}$ is a parameter which we assume is more stringent than the natural debt limit. We begin by positing that this asset is in zero aggregate net supply, as in Huggett (1993a), and traded among households in a competitive financial market.

Recursive formulation

We now restate the economy in recursive form. The dynamic programming version of the household problem described above is

$$V(a,y) = \max_{c,a'} \left\{ u(c) + \beta \sum_{y' \in Y} \pi(y',y) V(a',y') \right\} \tag{11.47}$$

subject to

$$c + a' = Ra + y$$

and

$$a' \geq -\underline{a},$$

where V denotes the value function, and the pair (a,y) is the individual state vector.

Let λ be the probability distribution of agents over states. Let $\bar{a}$ be the maximum asset holding in the economy, and for now assume that such upper bound exists. Let $A \equiv [-\underline{a}, \bar{a}]$ be the asset space. Let the state space S be the Cartesian product $A \times Y$. Let the σ-algebra Σ_s be defined as $B_A \otimes P(Y)$ where B_A is the Borel sigma-algebra on A and $P(Y)$ is the power set of Y. The space (S, Σ_s) is a measurable space. Let $\mathcal{S} = (\mathcal{A} \times \mathcal{Y})$ be the typical set of Σ_s. For any element of the sigma algebra $\mathcal{S} \in \Sigma_s$, $\lambda(\mathcal{S})$ is the measure of agents in the set $\mathcal{S}$.

How can we characterize the way individuals transit across states over time? We need a transition function. Let $Q((a,y), \mathcal{A} \times \mathcal{Y})$ be the (conditional) probability that an individual with current state (a,y) transits to the set $\mathcal{A} \times \mathcal{Y}$ next period. Formally, $Q : S \times \Sigma_s \to [0,1]$, and

$$Q((a,y), \mathcal{A} \times \mathcal{Y}) = \mathbb{I}_{\{a'(a,y) \in \mathcal{A}\}} \sum_{y' \in \mathcal{Y}} \pi(y',y) \tag{11.48}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function, and $a'(a,y)$ is the saving policy. This transition function can be used to construct a sequence of distributions as

$$\lambda_{n+1}(\mathcal{A} \times \mathcal{Y}) = \int_{A \times Y} Q((a,y), \mathcal{A} \times \mathcal{Y}) \, d\lambda_n. \tag{11.49}$$

We are now ready to proceed to the definition of a stationary, or steady-state, equilibrium.

Stationary equilibrium

A recursive stationary competitive equilibrium is: (a) value function $V : S \to \mathbb{R}$, (b) policy functions $a' : S \to \mathbb{R}$, and $c : S \to \mathbb{R}_+$, (c) an interest rate r^*, and (d) a stationary measure λ^* such that:

- Given r, the policy functions a' and c solve the household's problem (11.47) and V is the associated value function
- The asset market clears: $\int_{A\times Y} a'(a,y)d\lambda^* = 0$ at the interest rate r^*
- The goods market clears: $\int_{A\times Y} c(a,y)d\lambda^* = \sum_{i=1}^{N} y_i \Pi^*(y_i)$
- For all $(\mathcal{A}\times\mathcal{Y}) \in \Sigma_s$, the invariant probability measure λ^* is the fixed point of (11.49) and satisfies
$$\lambda^*(\mathcal{A}\times\mathcal{Y}) = \int_{A\times Y} Q((a,y),\mathcal{A}\times\mathcal{Y})\, d\lambda^*,$$
where Q is the transition function defined in (11.48).

In the stationary equilibrium, households optimize, markets clear, and the distribution of agents across states is invariant, i.e. this probability measure will reproduce itself permanently. Households, however, will exchange places and move upward and downward within this joint distribution of income and wealth, since they face different sequences of endowment shocks.

What guarantees that a steady-state exists and is unique? We start from existence.

Existence. Let us express the excess aggregate demand function for assets in financial markets as
$$A(r) = \int_{A\times Y} a'(a,y;r)d\lambda_r^*,$$
where we made explicit the dependence of the policy function and the stationary distribution on the interest rate r –think of r as a parameter for now. One of the conditions that yield existence is that $A(r)$ is continuous in r. $A(r)$ is continuous if both a' and λ_r^* are, in turn, continuous in r.

The theorem of the maximum implies that if u is continuous, $u' > 0$ and $u'' < 0$, the solution to the household problem is unique and the policy function $a'(a,\varepsilon;r)$ is continuous in r.

Stokey, Lucas and Prescott (Theorems 12.12 and 12.13) lay out conditions under which λ_r^* is continuous in r. Essentially, the stationary distribution must exist and be unique. We refer the interested reader to Stokey, Lucas and Prescott for the exact conditions needed and to Huggett (1993a) and Aiyagari (1994) for proofs that these assumptions are met in this environment. Most of these proofs are straightforward. We highlight two conditions that deserve more discussion, one for existence and one for uniqueness.

Existence of an invariant distribution requires compactness of the state space, i.e. a finite upper bound for wealth a. As explained, $\beta R < 1$ is a sufficient to obtain a finite upper bound for a, but R is endogenous so this restriction cannot be assumed. It must hold in equilibrium, and later in this section we show that it is true.

Uniqueness of the invariant distribution is guaranteed if the transition function $Q((a,y),\mathcal{A}\times\mathcal{Y})$ satisfies, for any given r, the "monotone mixing condition" (and a few regularity conditions). This condition states that there is a positive probability that a household moves from the bottom to the top of the asset space (and viceversa) in finite time. Namely, the economy

must have enough upward and downward mobility within the income and wealth distribution. To see why this condition is satisfied in our model, suppose the household starts from $(\bar{a}, y_{\max})$ and receives a long stream of the worst realization of the shock $y_{\min}$. If y follows a stationary (i.e., mean reverting) process, the household will keep decumulating wealth to smooth consumption until reaching some neighborhood of the lower bound. Symmetrically, suppose the household starts with $(-\underline{a}, y_{\min})$ and receives a long stream of the best shock $y_{\max}$. Knowing that sooner or later the shock will revert toward its mean, they will keep accumulating wealth until they reach some neighborhood of the upper bound.

Having proved that the $A(r)$ function is continuous, we need to argue that it crosses zero at least once at some finite value for r. For $\beta(1+r)=1$, we know from Section 11.3.5 that households will keep accumulating wealth without bound, so $A\left(\frac{1}{\beta}-1\right)=+\infty$. For low enough (possibly negative) values of r, every household would want to borrow. For example, it is clear that if $r=-1$, it is optimal to borrow up to the limit since one would never have to repay, and thus $A(-1)=-\underline{a}$. Since $A(r)$ is continuous, it will cross zero at least once and an equilibrium exists. This logic is represented graphically in Figure 11.6. Note that the equilibrium interest rate r^* will always lie below its complete market counterpart $1/\beta-1$. As a result, the condition $\beta R^*<1$ holds in equilibrium, which puts a limit to wealth accumulation and confirms that the asset space is bounded above.

A solution to the risk-free rate puzzle

The observation that the equilibrium interest rate is lower than under full insurance is important for asset pricing, as emphasized by Huggett (1993a). The representative agent model can only generate a high equity premium for very large levels of risk aversion (Mehra and Prescott, 1985) which, under CRRA utility, imply very low values for the intertemporal elasticity of substitution. In this situation, households have a very sharp desire to smooth consumption. When income has positive growth on average, as in the data, consumption smoothing dictates a strong desire to borrow against future income which pushes up the real rate well above observed values.

The cost of solving the equity premium puzzle with complete markets is, therefore, the "risk-free rate puzzle": the return on a risk-free bond is too high compared to the historical data. Huggett (1993a) showed that the presence of uninsurable idiosyncratic income risk can help solving the puzzle. As soon as one deviates from complete markets, the precautionary saving motive sets in (and a strong one if risk aversion is high) and, with it, a desire to save that pulls down the equilibrium interest rate on safe assets and better aligns it to the data.

Uniqueness. Having proved that the equilibrium exists, we ask: is the equilibrium unique? Uniqueness is guaranteed if the function $A(r)$ is monotonically increasing in r. This property is hard to prove in general because changes in r have both income and substitution effects on savings: the relative dominance between the two could switch at a certain level of assets, so $a'(a,\varepsilon;r)$ may not be monotone in r. Achdou, Han, Lasry, Lions, and Moll (2022) state a sufficient condition for CRRA utility in a continuous time version of this model. Let σ be

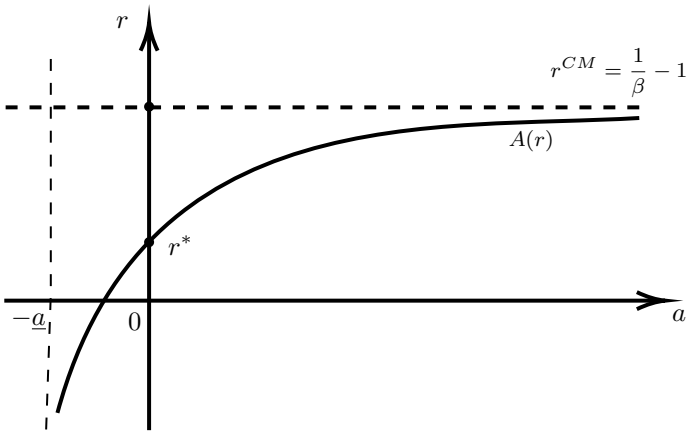


Figure 11.6: Equilibrium of the endowment economy where assets are in zero net supply.

Notes:$A(r)$ denotes aggregate asset holdings of the household sector as a function of the interest rate in the incomplete-market model. r^{CM} denotes the complete markets equilibrium real rate, and also the infinitely elastic demand for assets in the complete market model. r^* is the equilibrium real rate in the model with incomplete markets.

the coefficient of risk aversion. Recall that the smaller is σ the larger is the income effect and a higher r increases savings. Uniqueness is obtained if $\sigma < 1$.

Assets in positive supply

We conclude this section by extending this logic to an endowment economy where assets are not in zero net supply, but are one-period real bonds (or real balances of fiat money) issued by the government. This is the economy studied by Bewley (1980, 1983). Let B be this amount, fixed in steady-state. The government budget constraint is

$$rB = T \tag{11.50}$$

where T is the lump-sum tax (if $r > 0$) levied on households to finance interest payments. This is a new equilibrium condition determining T. The household budget constraint in (11.47) is modified as

$$c + a' = Ra + y - T,$$

and the equilibrium condition in the asset market, represented graphically in Figure 11.7, becomes

$$\int_{A\times Y} a'(a, y; r) d\lambda_r^* = B.$$

The analysis of equilibrium is virtually the same as in the previous section, with the only modification that we have an additional endogenous variable, T, and an additional equation, (11.50).

We conclude by noting that, in this economy, the government would have an incentive to provide more liquidity to households by increasing the level of debt B. Returning to Figure, as B increases, the economy moves closer to the full insurance outcome. In the economy with production that we analyze next, however, expanding government debt also entails a

cost, because it crowds out capital and production. Aiyagari and McGrattan (1998) analyze this trade-off in the determination of the optimal quantity of debt.

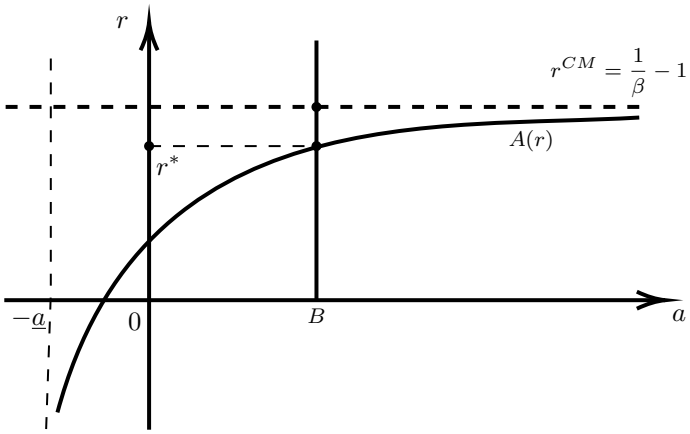


Figure 11.7: Equilibrium of the endowment economy where assets are in fixed positive supply.

Notes:$A(r)$ denotes aggregate asset holdings of the household sector as a function of the interest rate in the incomplete-market model. r^{CM} denotes the complete markets equilibrium real rate, and r^* the equilibrium real rate in the model with incomplete markets.

11.4.2 A production economy

We now follow Aiyagari (1994) and introduce production in this economy. We re-interpret y as efficiency units of labor (i.e., individual labor productivity) and assume that households supply labor inelastically, with each individual time endowment normalized to 1. The asset a represents financial claims to physical capital. Households supply labor and capital to a representative firm. The firm produces the final good with CRS production function $F(K, L)$, which uses capital and efficiency units of labor as inputs and satisfies $F_K > 0, F_L > 0, F_{KK} < 0, F_{LL} < 0$ as well as standard Inada conditions, $\lim_{K\to\infty} F_K = 0$ and $\lim_{K\to 0} F_K = +\infty$. Physical capital depreciates geometrically at rate $\delta \in (0, 1)$. Firms act competitively by maximizing profits taking prices as given. The final numeraire good can be used for consumption and investment, and is traded in a competitive good market. The labor market and capital market are also competitive and clear, respectively, at the wage rate w (per efficiency unit) and interest rate r.

The household budget constraint in problem (11.47) becomes

$$c + a' = Ra + wy$$

and everything else in the household problem is unchanged. We now state the new definition of equilibrium.

A recursive stationary competitive equilibrium is: (a) value function $V : S \to \mathbb{R}$, (b) policy functions $a' : S \to \mathbb{R}$, and $c : S \to \mathbb{R}_+$, (c) firm choices L and K, (d) prices r^* and w^*, and (e) a stationary measure λ^* such that:

- Given (r, w), the policy functions a' and c solve the household's problem (11.47) and V is the associated value function

- Given (r, w), the firm chooses optimally its capital stock K and its labor input L, i.e., $r + \delta = F_K(K, L)$ and $w = F_L(K, L)$

- The labor market clears: $L = \sum_{i=1}^{N} y_i \Pi^* (y_i)$ at the wage w^*

- The asset market clears: $\int_{A \times Y} a'(a, y) d\lambda^* = K$ at the interest rate r^*

- The goods market clears: $\int_{A \times Y} c(a, y) d\lambda^* + \delta K = F(K, L)$

- For all $(\mathcal{A} \times \mathcal{Y}) \in \Sigma_s$, the invariant probability measure λ^* satisfies

$$\lambda^* (\mathcal{A} \times \mathcal{Y}) = \int_{A \times Y} Q ((a, y), \mathcal{A} \times \mathcal{Y}) \, d\lambda^*,$$

Production changes the shape of the aggregate demand for capital from firms. To characterize it, we need to consider the firm's problem. From the optimal choice of the firm, we obtain $K(r)$ implicitly from $F_K(K, L) = r + \delta$. It is immediate to see that for $r = -\delta$, then $K \to +\infty$, while for $r \to +\infty$, $K \to 0$, given our assumptions on F, in particular the Inada conditions. Thus, firms' demand for capital is a continuous, strictly decreasing function of the interest rate r.[27] Figure 11.8 illustrates the equilibrium in this model. Existence follows from the same conditions discussed above.

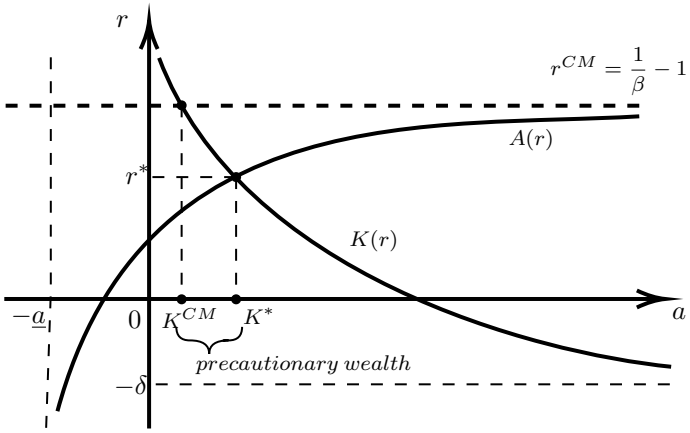


Figure 11.8: Equilibrium of the production economy.

Notes: $A(r)$ denotes aggregate asset holdings of the household sector as a function of the interest rate in the incomplete-market model. r^{CM} and K^{CM} denote, respectively, the complete markets equilibrium real rate and capital stock. r^* and K^* are their incomplete markets counterparts. The difference between K^* and K^{CM} is the equilibrium amount of precautionary wealth.

[27] For example, if $F(K, L) = K^\alpha L^{1-\alpha}$, then $K(r) = \left(\frac{\alpha L}{\delta + r}\right)^{\frac{1}{1-\alpha}}$.

Quantitative analysis of the model.

Aiyagari (1994) calibrated the model to the U.S. economy. One of the key ingredients of the model, the stochastic process of earnings shocks (in particular, their persistence and volatility), is estimated from longitudinal micro data on individual labor income. Aiyagari reached three important conclusions from his quantitative analysis, all results that have shaped the literature for decades.

First, he argued that this incomplete market structure is quite effective at insuring income risk. By saving, dissaving, and borrowing (i.e., through self-insurance), an individual can cut consumption variability by about half compared with autarky. Second, he pointed out that one can use the model to quantify the amount of aggregate wealth held by households because of precautionary reasons. The capital stock in complete markets is the value K^{CM} where the aggregate demand for capital by firms crosses the infinitely elastic supply of capital of the representative consumer at $r^{CM} = 1/\beta - 1$. The difference between K^* and K^{CM} is therefore the additional precautionary stock of capital. He concluded that, in his baseline calibration, the precautionary motive augments the aggregate saving rate by 3 percentage points, but calibrations featuring higher risk aversion or more income volatility can raise the aggregate saving rate by 10 percentage points or more. Third, he observed that the model can generate an equilibrium wealth distribution that is more positively skewed (mean > median) and more dispersed than the income distribution, as in the data. Compared to its empirical counterpart, though, the model's wealth distribution features too much wealth accumulation at the bottom and too little wealth concentration at the top. Chapter 21 describes the progress made in the literature on wealth inequality since Aiyagari's observation.

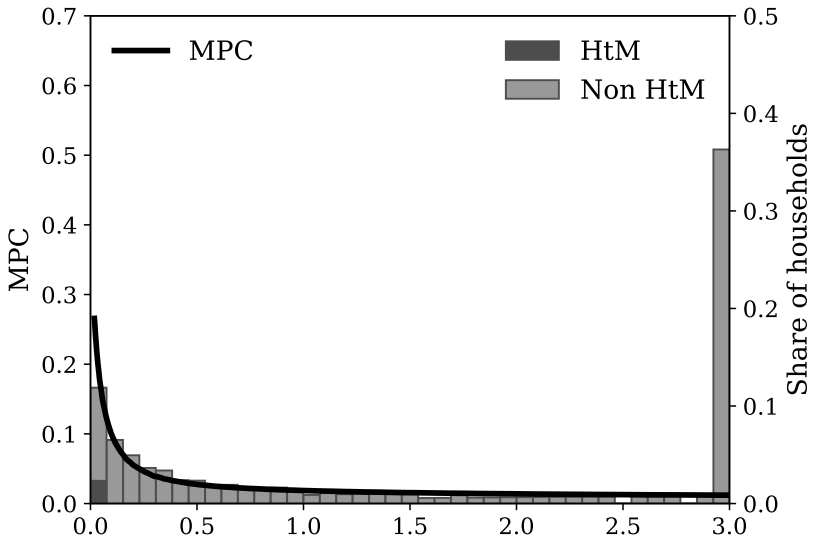


Figure 11.9: MPC, i.e. the slope of the consumption decision rule (curve), as a function of wealth jointly with the distribution of wealth (bars).

Notes: The dark bar near zero represents the share of constrained (or hand-to-mouth, HtM) households in the calibrated model. This figure is reproduced from Kaplan and Violante (2022)

The marginal propensity to consume: models vs data

In Section 11.3.1 we showed that, under the permanent income hypothesis (PIH), the marginal propensity to consume (MPC) out of wealth and unanticipated transitory changes in income equals $1 - R^{-1}(\beta R)^{\frac{1}{\sigma}}$. In the equilibrium of a representative agent economy, where $\beta R^* = 1$, the MPC therefore roughly equals the real interest rate, or around 1% on a quarterly basis. In addition, according to the PIH, the MPC out of anticipated income changes is zero.

There is a vast body of work that estimates the MPC out of (more or less anticipated) transitory changes in income using various approaches. Some of the most convincing evidence comes from the observation that when households receives fiscal stimulus payments from the government (like in the last three recessions, for example), they spend on average around 15-30% of this transfer on nondurable goods in the first quarter after receipt (e.g., Parker, Souleles, Johnson, and McClelland, 2013). These estimates indicate that, empirically, the MPC might be up to 20 times as large in the data than what implied by the naive PIH.

The heterogeneous-agent incomplete-market model we analyzed has the potential to match the data better. As discussed in Section 11.3.4, for low enough wealth levels the slope of the consumption function can be much steeper than r. The key question is: in the equilibrium of a plausibly calibrated model, what is the share of agents with wealth holdings in that range? The answer is not many for two reasons. First, in order for the model to match the aggregate wealth level observed in the U.S., the model discount factor has to be high. Second, precisely because being on a constraint induces excessive consumption fluctuations that are costly in terms of welfare, optimization leads households to save and keep away from their credit limit. Overall, the quarterly MPC can be boosted up to 5% or so, but no more. Figure 11.9, reproduced from Kaplan and Violante (2022) illustrates this result.

The literature has studied how to modify the baseline model to generate levels of average MPC more in line with the data. We refer to Kaplan and Violante (2022) for a comprehensive survey, and here only discuss two of them. First, one can introduce heterogeneity in discount factors across consumers. Households with low discount factors display steeper consumption functions (recall our comparative statics of Section 11.3.4). In addition, because of impatience, they tend to dissave and thus hold small amounts of wealth. With enough households of this type, the average MPC of the economy can be large. Second, one can take the view that the relevant notion of wealth for short-term consumption smoothing is *liquid wealth*, e.g., cash and bank deposits but not housing or retirement accounts. Empirically, liquid wealth is only a small fraction of total wealth. A model with two types of assets, a liquid one and an illiquid one which can only be accessed by paying a transaction cost, is able to yield averages MPC in line with the data. Interestingly, this type of models features *wealthy hand-to-mouth* consumers, i.e. consumers who do have substantial wealth locked in high-return illiquid assets, but hold small amounts of liquid wealth and thus are highly responsive to transitory income changes (Kaplan and Violante, 2014; Kaplan, Violante, and Weidner, 2014).

Chapter 12

Labor supply

Richard Rogerson and Giovanni L. Violante

12.1 Introduction

Hours worked are a fundamental input in the aggregate production function and, consequently, a crucial determinant of an economy's output per capita. Macroeconomists conceptualize the total hours worked at a given point in time as the equilibrium outcome where demand meets supply in the labor market. In this chapter, we examine the factors influencing labor supply.

The macroeconomic study of labor supply is motivated by a set of stylized facts regarding the variation in hours worked: across countries, over time both in the long run at the low frequency of economic growth and in the short run at the business cycle frequency, throughout an individual's life cycle, and across demographic groups, such as men and women, young and old, and skilled and unskilled workers. Understanding these variations hinges on three key elements. First, accurately identifying the incentives and disincentives that influence individuals' decisions to work. Second, determining the sensitivity of individuals to these incentives, measured by labor supply elasticities. Third, appropriately aggregating these elasticities to derive the macroeconomic patterns of hours worked. Each of these steps involves subtle considerations and is the focus of an extensive body of research.

The chapter is organized as follows. Section 12.2 summarizes some key facts about variation in hours of work. Section 12.3 introduces a benchmark static model of labor supply and introduces several key elasticity concepts that are relevant for understanding how labor supply responds to changes in the economic environment. It also describes several enrichments of the static model. Section 12.4 extends our results derived for the benchmark static model of Section 12.3 to a dynamic stochastic life-cycle model. There, we also discuss the empirical literature that estimates labor supply elasticities with micro data, including a discussion of how optimization frictions can affect estimation. Section 12.5 derives the restrictions that balanced growth imposes on preferences over consumption and hours worked. Section 12.6 discusses how macroeconomists have used representative agent models to tackle variation in hours worked across space and over time. First, we explain how differences in the level of taxes across countries can account for the large observed differences in hours worked. Second, we illustrate the role of the Frisch elasticity in fluctuations of hours worked over the business cycle. The final two sections of the chapter describe extensions that have

important implications for labor supply elasticities and their measurement. Section 12.7 generalizes the dynamic model of Section 12.4 to allow for human capital accumulation, thereby introducing dynamic returns to current labor supply. Section 12.8 introduces models that explicitly consider labor supply along both the extensive margin (employment) and intensive margin (hours when employed). We highlight how these additional features have important implications for labor supply elasticities.[1]

A key message that we want the reader to take away from this chapter is that a relatively simple and parsimonious model of labor supply can help us understand many of the patterns found in the data on aggregate labor market outcomes. For this reason it is important for macroeconomic models to explicitly incorporate labor supply. Importantly, this message should not be interpreted to imply that all labor market outcomes reflect desired labor supply by workers. Modern models of unemployment stress the possibility that frictions in the labor market create a wedge between labor market outcomes and desired labor supply. Frictional labor markets are discussed in Chapter 20.

12.2 Facts about hours of work

In this section we present some basic facts about hours of work, focusing on differences in hours of work in several contexts. The first subsection documents differences across countries, and decomposes these differences along the intensive (hours per worker) and extensive (fraction of people working) margins. The second subsection examines differences across time, and in particular secular trends and business cycle fluctuations. The third subsection documents differences across demographic groups within the U.S.

12.2.1 Differences across countries

One of the most basic statistics for macroeconomists is aggregate hours of work normalized by some measure of population.

Table 12.1: Hours of work per person aged 15+ relative to the U.S. 2015-2019

<0.75	**(.75,.85)**	**(.85,.95)**	**>.95**
Italy (0.69)	Finland (0.77)	UK (0.85)	Canada (0.96)
France (0.70)	Austria (0.79)	Sweden (0.90)	Australia (0.98)
Belgium (0.72)	Norway (0.80)	Ireland (0.91)	U.S. (1.00)
Greece (0.73)	Netherlands (0.82)	Japan (0.91)	New Zealand (1.07)
Denmark (0.74)	Portugal (0.85)	Switzerland (0.93)	Korea (1.12)
Germany (0.74)			
Spain (0.75)			

[1]To the student who wants to deepen their understanding of labor supply, we recommend the classic handbook chapter by Blundell and MaCurdy (1999) and the more recent survey articles by Keane (2011), Keane and Rogerson (2012, 2015), and Rogerson (2024).

Table 12.2: Labor supply along the intensive and extensive margin 2015-2019

Panel A: Extensive Margin: Employment to Population (%)				
<50	**(50,55)**	**(55,60)**	**(60,65)**	**>65**
Greece (40.9)	Belgium (50.0)	Ireland (57.7)	U.S. (60.1)	Switzerland (65.1)
Italy (44.1)	France (50.5)	Austria (57.9)	UK (60.4)	New Zealand (66.9)
Spain (48.6)	Portugal (52.3)	Denmark (58.3)	Korea (60.7)	Sweden (67.6)
	Finland (54.4)	Germany (58.8)	Australia (61.7)	
		Japan (59.0)	Canada (61.8)	
			Netherlands (62.3)	
Panel B: Intensive Margin: Annual Hours Worked per Employed Person				
<1500	**(1500,1650)**	**(1650,1750)**	**(1750,1850)**	**> 1850**
Germany (1388)	Austria (1502)	Spain (1693)	New Zealand (1761)	Greece (1941)
Denmark (1395)	France (1516)	Japan (1693)	U.S. (1825)	Korea (2026)
Norway (1423)	UK (1535)	Canada (1699)		
Netherlands (1435)	Finland (1549)	Italy (1718)		
Sweden (1466)	Switzerland (1563)	Ireland (1720)		
	Belgium (1577)	Portugal (1736)		
		Australia (1737)		

Table 12.1, adapted from Rogerson (2024), compares aggregate hours of work per person aged 15 and older across 23 OECD economies for the period 2015-2019. To facilitate comparisons all values are reported relative to the U.S. The value for the U.S. is 1096 hours per year. These differences in aggregate hours per person reflect differences both in the hours per person employed (the *intensive* margin) and the fraction of the population that is employed (the *extensive* margin). Table 12.2 presents differences along each of the two margins for the same countries and time period as Table 12.1. As a reference point for interpreting the values in Panel B, we note that an individual who works 40 hours a week for 50 weeks will have 2000 annual hours.

12.2.2 Differences over time

Macroeconomists are also very interested in the time series changes in aggregate hours of work. Figure 12.1 plots the log of aggregate hours per person aged 15 and older for the G7 economies over the period 1950-2019. Two properties are worth noting. First, most of the countries experience a very large decline in aggregate hours per person over this period, with several countries experiencing declines of roughly 30 percent. The U.S. is somewhat of an exception, with hours in 2019 only marginally lower than in 1950 and with no statistically significant trend. Second, the large differences that we documented in Table 12.1 for the period 2015-2019 are not a constant feature of the data. Although the U.S. has one of the highest values for aggregate hours per person in 2019, it actually has the lowest value among these countries in 1950.

It is also of interest to look at the separate time series for the intensive and extensive margins. Figure 12.2 shows these series for the G7 economies over the same time period.

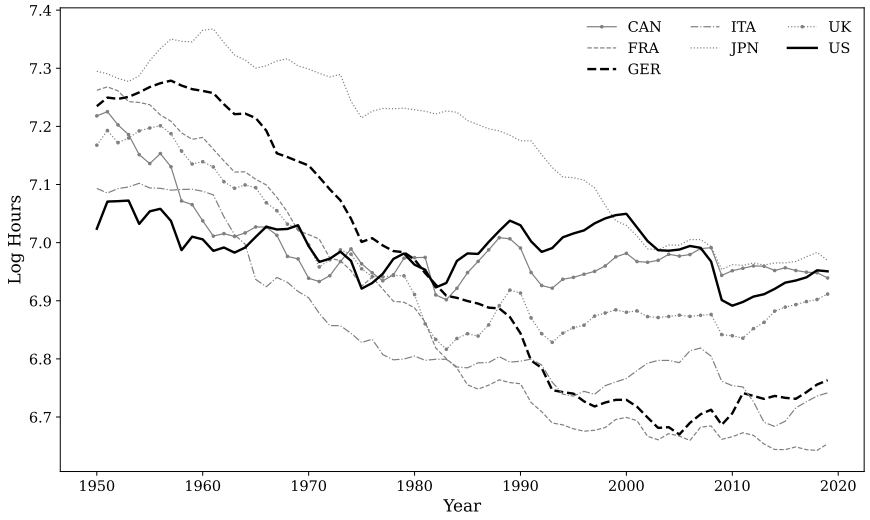


Figure 12.1: Log average annual hours worked per person for the G7 countries from 1950-2019

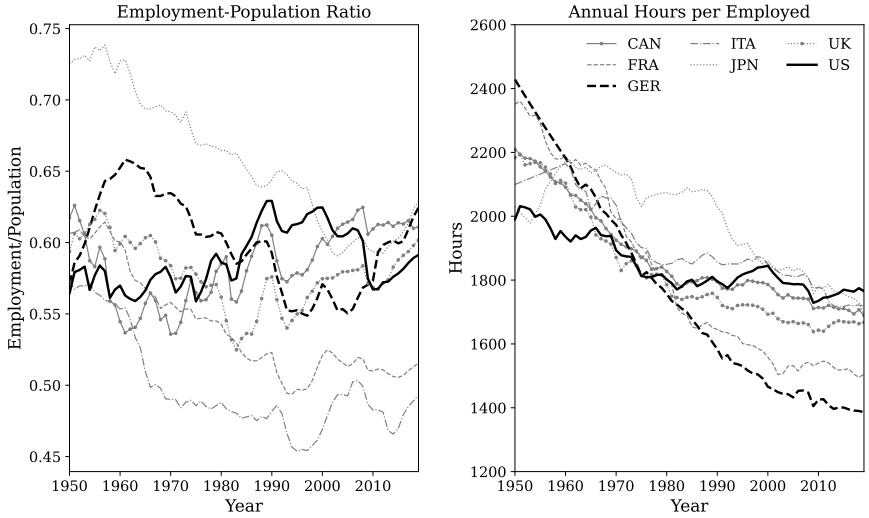


Figure 12.2: Intensive and extensive margins of labor supply for G7 countries from 1950 to 2019.

Notes: Left panel: employment to population ratio. Right panel: annual hours per employed person.

There are some notable differences between these two figures. Regarding the intensive margin, we see that all countries experience relatively large secular declines, though there is a large range in the magnitude of the decline. Germany experiences a reduction of roughly 40 percent, while the U.S. experiences a decline of roughly 10 percent. The country level experiences for changes along the extensive margin are much more varied. Some countries witness a significant secular decline, while others exhibit relatively little secular change, and still others feature a secular increase.

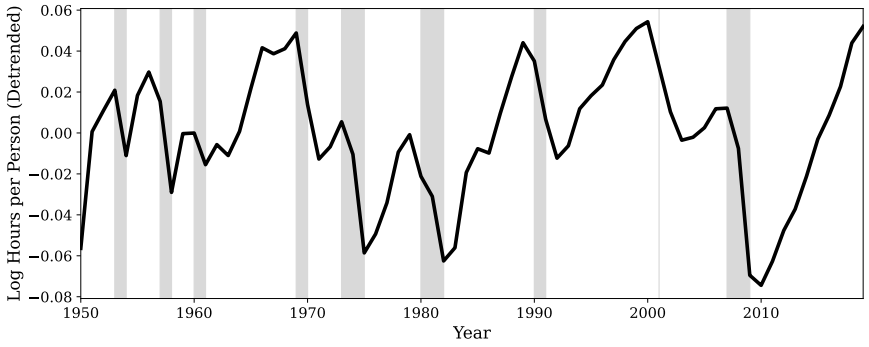


Figure 12.3: Business cycle fluctuations worked per person for the U.S. from 1950-2019

.

Notes: Calculated as the residual of a regression of log average annual hours worked per person on a quartic polynomial in time.

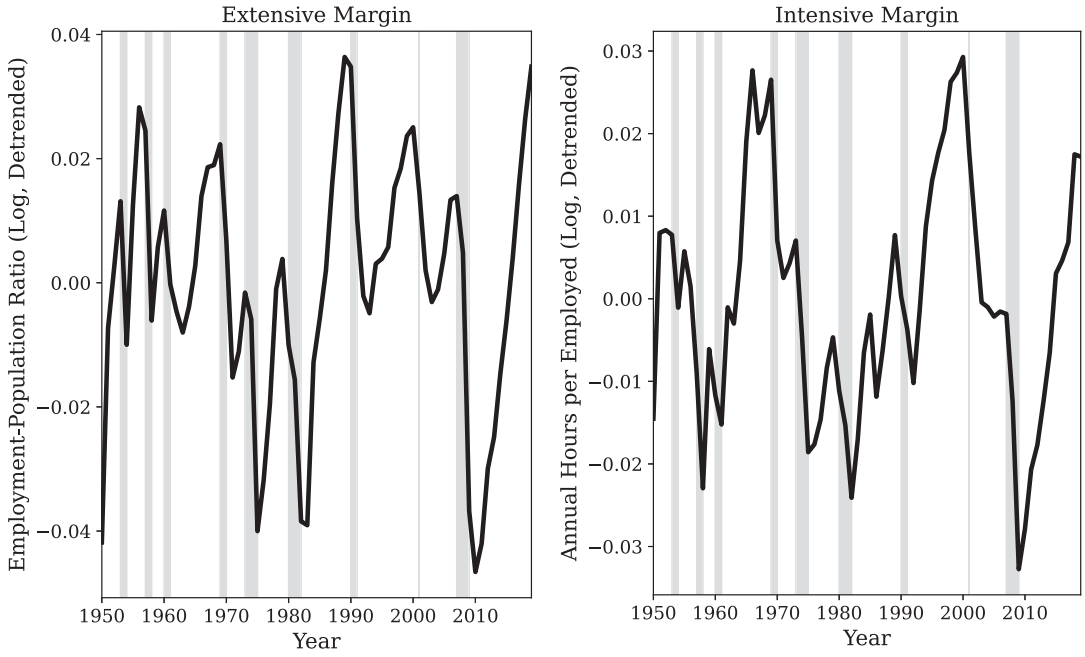


Figure 12.4: Business cycle fluctuations in labor supply on the intensive (left panel) and extensive (right panel) margins for the U.S. from 1950-2019.

Notes: The business cycle variation in the intensive margin of labor supply is calculated as the residual of a regression of log annual hours worked per employed on a quartic polynomial in time. The business cycle variation in the extensive margin of labor supply is calculated as the residual of a regression of log employment to population ratio on a quartic polynomial in time.

When examining time series for hours worked, macroeconomists are often interested in distinguishing between trend behavior and deviations from trend, typically associated with business cycle fluctuations.

There are various methods used to perform this decomposition. Here we simply use a quartic polynomial in time to identify the trend and define the business cycle to be deviations from this trend. Figure 12.3 displays the business cycle component of aggregate hours for the U.S. economy, with grey shaded areas indicating NBER recession dates. This figure displays the well-known fact that aggregate hours display large fluctuations over the business cycle, with the movements from peak to trough exceeding ten percentage points in many instances. One might also be interested in the nature of business cycle fluctuations along the intensive and extensive margins separately. The two panels of Figure 12.4 show this for the U.S. economy. The left panel displays fluctuations along the extensive margin while the right panel shows fluctuations along the intensive margin. Both exhibit relatively large fluctuations, though the movements along the extensive margin tend to be somewhat larger.

12.2.3 Differences across demographic groups

To this point we have focused on purely aggregate measures of labor supply. Another important feature of the data is that hours of work vary quite dramatically across demographic groups within a country. To illustrate this we rely on data from the American Time Use Survey for the year 2023. This survey produces data on average time devoted to work per day broken down by various demographic characteristics. Table 12.3 shows how time devoted to work varies by gender and age group.

Table 12.3: Hours of work per day

Age	**All**	**Male**	**Female**
$15-19$	1.27	1.24	1.29
$20-24$	3.67	3.81	3.52
$25-34$	5.02	5.90	4.13
$35-44$	5.04	5.79	4.29
$45-54$	4.79	5.45	4.15
$55-64$	4.02	4.73	3.35
$65-74$	1.42	1.88	1.02
75+	0.37	0.55	0.23
Total	3.56	4.17	2.98

For both males and females we see a very pronounced hump shape for time devoted to work over the life cycle, with hours peaking during the 25-54 age range and falling off quite substantially at younger and older ages. Although both males and females display the same hump-shaped pattern, the life cycle profile for female time devoted to work is substantially lower than the profile for males, with the lone exception of work among 15-19 year olds. Gender differences have also changed dramatically over time. Figure 12.5 shows the evolution of the employment to population ratio for males and females in the U.S. since

1950. At the same time that the male employment to population ratio has declined by more than ten percentage points the female employment to population ratio has increased by roughly 20 percentage points. These large changes in gender employment differences are a pervasive feature of the data.

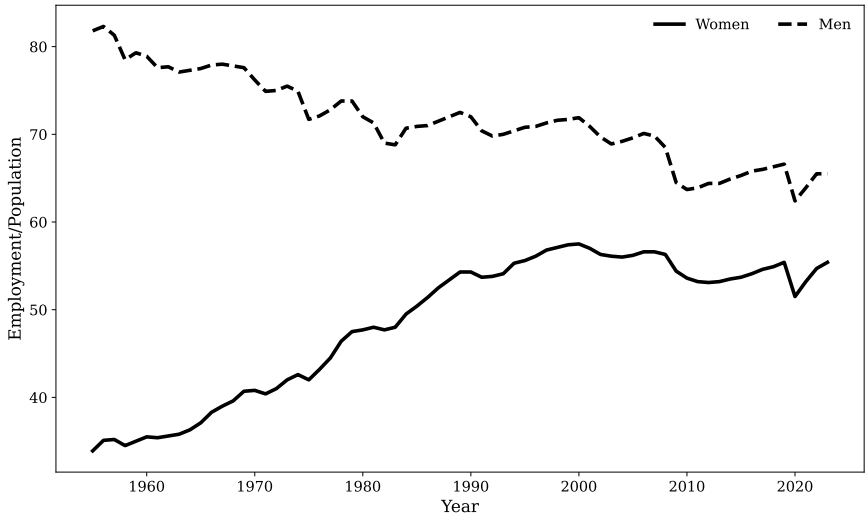


Figure 12.5: Employment to population by gender for the U.S. by gender from 1950-2019.

In addition to large differences in hours of work by age and gender, there are also large differences by educational attainment. Table 12.4 shows average time devoted to work per day by educational attainment for five educational attainment groups: less than high school (<HS), exactly high school (HS), some college but not a college degree (SC), a four year college degree but nothing further (C) and more than a college degree (>C).

Table 12.4: Hours of work per day by education

<HS	HS	SC	C	>C
2.89	3.39	3.90	4.08	4.17

This table shows that average time devoted to work increases substantially with educational attainment, with college educated workers devoting roughly 20% more time to work relative to high school educated workers.

12.2.4 Summary

In this section we have documented that there are large differences in hours of work across countries, across time, and across demographic groups. Understanding the forces that account for these differences is an important goal of economics research. One key question in this context concerns the extent to which these differences in hours of work reflect different choices by individuals, and if so, what factors account for these different choices. Answering

this question requires that one think about the forces that shape individual labor supply decisions. The goal of this chapter is to introduce the student to the basic economics of labor supply.

12.3 The theory of labor supply: static models

Much of this chapter is devoted to understanding the forces that shape labor supply in dynamic stochastic settings. But because the same forces that shape labor supply in simpler settings will also play an important role in dynamic stochastic environments, it is useful to begin our analysis of labor supply by considering a simple textbook model of labor supply in a static and deterministic setting. This setting is sufficient to introduce several elasticity concepts that are important in understanding how labor supply responds to various changes in the economic environment.

12.3.1 A benchmark model of labor supply

In this section we study a standard textbook model of individual labor supply and examine some of the forces that shape optimal labor supply. We consider an individual with preferences over consumption (c) and hours worked (h). (In earlier chapters, we used ℓ as working hours; this chapter uses a different notation.) While the discussion in this section can be generalized to a large class of preferences, to facilitate exposition we focus on a special case which is commonly used in the macroeconomics literature. Specifically, we assume that preferences are separable between consumption and hours of work:

$$U(c,h) = u(c) - v(h).$$

with the functions $u(c)$ and $v(h)$ of the following form:

$$u(c) = \frac{1}{1-\sigma}c^{1-\sigma} \text{ and } \; v(h) = \frac{\psi}{1+1/\gamma}h^{1+1/\gamma},$$

where ψ, σ, and γ are all positive constants. As we show later in this chapter, this relatively parsimonious specification of preferences leads to a theory of labor supply that helps us understand many features documented in the previous section.

The price of consumption is normalized to one, the individual faces a wage rate of w, and has non-labor income equal to I. The individual thus solves the following problem:

$$\max_{c,h} \; U(c,h)$$

subject to

$$c = wh + I,\; c \geq 0, \text{ and } h \geq 0.$$

Because the marginal disutility of work is equal to zero when $h = 0$, the individual will always choose $h > 0$. Substituting the budget equation into the utility function one obtains the first order condition:

$$\frac{\psi h^{1/\gamma}}{(wh+I)^{-\sigma}} = w. \tag{12.1}$$

This equation has the interpretation that the individual chooses hours of work h so that the marginal rate of substitution between consumption and working time is equal to the level of the real wage w, i.e., the price of time relative to consumption.

Individuals differ both in the wage rate that they face and their level of non-labor income. It is of interest to ask what our model implies for how these differences will affect labor supply. Economists typically find it useful to summarize these effects as elasticities. If we write the solution to the optimal labor supply problem as $h(w, I)$, then the wage and income elasticities, denoted by $\varepsilon^M_{h,w}$ and $\varepsilon_{h,I}$ can be written as:

$$\varepsilon^M_{h,w} = \frac{w}{h} h_w(w, I)$$

and

$$\varepsilon_{h,I} = \frac{I}{h} h_I(w, I).$$

Using the implicit function theorem and equation (12.1) and carrying out some simple algebra yields the following two expressions:

$$\varepsilon_{h,I} = -\frac{\sigma}{(1/\gamma) + \sigma(1 - s_I)} s_I \tag{12.2}$$

and

$$\varepsilon^M_{h,w} = \frac{1 - \sigma(1 - s_I)}{(1/\gamma) + \sigma(1 - s_I)}, \tag{12.3}$$

where $s_I = I/(wh + I)$ is the share of non-labor income in total income.

Note that there is an intimate connection between curvature parameters in the utility function and these two labor supply elasticities. This implies that knowledge of labor supply elasticities provide information about the parameters of the utility function. We begin by discussing the income elasticity. Given that both σ and γ are positive and $s_I \leq 1$, we see that the income elasticity is always negative if s_I is positive.[2] The intuition for this result is straightforward. Higher non-labor income implies a higher level of consumption for any given level of hours. Because the marginal utility of consumption is decreasing, higher consumption decreases the marginal benefit of working and thus causes the individual to reduce hours of work.

Next we turn to the wage elasticity. One of the most fundamental questions in the context of labor supply is how does the optimal choice of h respond to an increase in w? Looking at equation (12.3), one sees that the sign of the denominator is necessarily positive, but that the sign of the numerator could be positive or negative depending upon the values of σ and s_I. The reason for this ambiguous prediction is that an increase in w has two opposing effects on the optimal choice of h. On the one hand, an increase in w increases the reward to additional work at the margin, thereby creating an incentive for additional work. This is what economists call the **substitution effect**. On the other hand, holding hours fixed, an increase in w raises the consumption of the individual, which makes additional

[2] If $s_I = 0$ then the income elasticity is equal to zero. But importantly, this does not imply that the effect of income on hours is zero. The same algebraic procedure that one uses to solve for the income elasticity implies that $h_I(w, I) = -\sigma(1 - s_I)/[(1/\gamma) + \sigma(1 - s_I)]$. In particular, when s_I equals zero we see that $h_I(w, I) = -\sigma/[(1/\gamma) + \sigma]$, which is necessarily negative.

consumption less valuable at the margin, thereby creating an incentive for less work. This is what economists call the **income effect**.

It is useful to focus on the special case in which non-labor income is equal to zero. In this case we have:

$$\varepsilon_{h,w}^{M} = \frac{1-\sigma}{(1/\gamma)+\sigma}.$$

It follows that the wage elasticity is strictly positive if σ is less than one, strictly negative if σ is greater than one and exactly equal to zero if σ is equal to one. Put somewhat differently, when $\sigma < 1$ the substitution effect dominates the income effect and when $\sigma > 1$ the income effect dominates the substitution effect. When $\sigma = 1$ changes in the wage have no effect on optimal labor supply. Importantly, when $\sigma = 1$ neither the income nor the substitution effect is equal to zero; rather, they are exactly offsetting. Specifically, the magnitude of each effect is increasing in the value of γ.

The elasticity of hours with respect to the wage that we have calculated is referred to as the **Marshallian elasticity**, and explains why we have included a superscript M. Our preceding discussion suggests that this elasticity can be thought of as the sum of two pieces, one reflecting the substitution effect and the other reflecting the income effect. Holding hours fixed, the additional income associated with a marginal increase in w is equal to level of hours worked, so that the size of the income effect is given by $h \cdot h_I(w, I)$. Given that the total effect on hours is given by $h_w(w, I)$, we can compute the magnitude of the substitution effect as the difference between the total effect of the wage change and the income effect associated with the wage change:

$$\text{substitution effect} = h_w(w, I) - h \cdot h_I(w, I).$$

To express this in terms of elasticities we multiply both sides by w/h and rearrange to obtain:

$$\varepsilon_{h,w}^{H} = \varepsilon_{h,w}^{M} - (1 - s_I)\varepsilon_{h,I},$$

where $\varepsilon_{h,w}^{H}$ is what economists refer to as the **Hicksian elasticity**. This elasticity reflects the change in hours worked in response to a change in the wage rate holding utility constant. Substituting from our earlier expressions we obtain:

$$\varepsilon_{h,w}^{H} = \frac{1}{(1/\gamma)+\sigma(1-s_I)}. \tag{12.4}$$

A key result is that the Hicksian elasticity is always positive. Moreover, it is always larger than the Marshallian elasticity.

It is perhaps useful to provide a concrete example to illustrate the usefulness of the different elasticities that we have introduced. An important policy question in which labor supply plays a key role is the effects of tax and transfer programs. To pursue this we introduce two policy parameters into the individual labor supply problem that we have been studying: a proportional tax on labor income at rate τ and a lump-sum transfer payment denoted by T. The individual's labor supply problem now becomes:

$$\max_{c,h} \; U(c, h)$$

subject to

$$c = (1-\tau)wh + I = (1-\tau)wh + y + T,\ c \geq 0, \text{ and } h \geq 0,$$

where y is non-labor income apart from the government transfer payment T. If we are solely interested in the effect of changes in the proportional tax rate on hours of work, we see that a change in $(1-\tau)$ is equivalent to a decrease in w, so that the effect of interest is given by the Marshallian elasticity. If we are solely interested in the effect of a change in transfer payments on hours of work, we see that a change in T is equivalent to a change in I, so that the effect of interest is given by the income elasticity. Changing tax rates and transfer payments individually have implications for the government budget. If one wants to study changes in tax and transfer policies that are budget neutral then one could focus on the case in which the transfer payment is equal to the amount of revenue raised by the tax, i.e., $T = \tau wh$. In this case there is no income effect associated with the change in τ and the appropriate elasticity for assessing the impact on hours is the Hicksian elasticity. In a highly cited paper, Prescott (2004) essentially used this framework to study the effect of differences in tax and transfer systems between the U.S. and several Western European countries on hours of work. He found that the effects were large. We examine this in more detail in Section 12.6.1.

12.3.2 Richer versions of the static labor supply model

In this section we consider three extensions to illustrate additional channels that have also been shown to play an important role in shaping labor supply.

Home production

In our previous analysis, all of the work that individuals did occurred in the market at the going wage of w, and all of their consumption was purchased in the market. In reality, individuals engage in many work activities outside of the market and there are several components of consumption that individuals can produce for themselves without purchasing them in the market. Some prominent examples include cooking and cleaning services, yard work, and child care. Economists use the term home production to describe the time that individuals devote to these activities. Classic references include Becker (1965) and Gronau (1977). We now generalize our previous analysis to allow for home production.

To do this we introduce the notion of the home production function and consider two alternative types of time spent working: h_m will be time spent working in the market and h_h will be time spent on home production. As before, time spent working in the market will generate income according to the market wage rate w. Time spent working at home will now generate home production, which we capture by the production function $y_h = A_h h_h$, where A_h captures the productivity of time spent in home production.[3] We now assume that the total consumption of an individual is given by the composite of market consumption denoted

[3]More generally, one could generalize this to allow for diminishing marginal product by writing $y_h = A_h f(h_h)$, where the function f is strictly concave. One could also add capital as an additional input. We discuss the significance of this generalization at the end of the subsection.

by c_m and consumption of home production, denoted by c_h:

$$c = g(c_m, c_h).$$

Our analysis will follow much of the literature and assume that the function g is given by a constant elasticity of substitution function:

$$g(c_m, c_h) = [\theta c_m^{\eta} + (1-\theta)c_h^{\eta}]^{\frac{1}{\eta}}$$

with $0 < \eta < 1$, indicating that home and market consumption are substitutes. We assume that the disutility of work depends only on the sum of market work and home production time, though one could generalize this.

The individual now solves the following optimization problem:

$$\max_{h_m, h_h} \; u\left(g(h_m w + I, A h_h)\right) - v(h_m + h_h)$$

subject to

$$h_m \geq 0 \text{ and } h_h \geq 0,$$

where the functions u and v are as before, though for much of our analysis the specific functional forms for these two functions will not play any role. From this maximization problem we obtain the following two first order conditions:

$$\begin{aligned} h_m &: u' g_1 w = v' \\ h_h &: u' g_2 A = v'. \end{aligned}$$

Dividing the two first order conditions by each other gives:

$$\frac{g_2}{g_1} = \frac{w}{A}. \tag{12.5}$$

This equation has an intuitive economic interpretation: it states that the marginal rate of substitution between home and market consumption is equal to the marginal rate of transformation between home and market consumption. For our purposes we want to focus on one particular implication. Specifically, assuming g is a CES aggregator and focusing on the special case in which $I = 0$ we obtain the following expression for the left hand side of equation (12.5):

$$\frac{g_2}{g_1} = \frac{1-\theta}{\theta}\left[\frac{c_h}{c_m}\right]^{\eta-1} = \frac{1-\theta}{\theta}\left[\frac{A h_h}{w h_m}\right]^{\eta-1}.$$

Substituting this into equation (12.5) and rearranging terms yields:

$$\frac{h_h}{h_m} = \left[\frac{1-\theta}{\theta}\right]^{\frac{1}{\eta-1}}\left[\frac{w}{A}\right]^{\frac{\eta}{\eta-1}}. \tag{12.6}$$

If $0 < \eta < 1$ then the exponent on w/A is negative and we have that the ratio of home to market work is negatively related to the ratio of the market wage to the level of home productivity.

The key message from this extension is that technical change that affects the productivity of time spent in home production can have important effects on labor supply to the market sector. Greenwood, Seshadri, and Yorukoglu (2005) used a generalization of this model to argue that technological changes in home production played a major role in the rise of female labor force participation observed in the first part of the 20th century. Their argument was that the invention of home capital goods like the washing machine served to greatly reduce the required labor input for home production, and thereby lowered A, the marginal product of labor in home production. In accordance with equation (12.6), this leads to a reallocation of time from the home sector to the market sector.

Household labor supply

The second extension considers labor supply from a household perspective when we explicitly consider households that consist of multiple individuals. The motivation for this is straightforward: the vast majority of total hours of work in modern economies is accounted for by individuals who live in households with multiple members. The key point we make here is that when looking at labor supply at the individual level, it is not only the market wage of that individual that matters, but also the market wages of other individuals in their household. Here we make this point in the context of the most commonly studied situation, that of two member households.

Studying optimal labor supply in multi-member households necessarily raises the question of how to specify the objectives and interactions of individuals within the household. Do they behave strategically with regard to each other? Do they behave cooperatively? The literature on family labor supply has considered different possibilities, but here we will focus on a common and particularly tractable specification, in which we assume that all family members share a common objective function. In the literature this is referred to as the unitary household model.[4] It eliminates any potential for strategic considerations between household members since there is no disagreement over which outcomes are best. While a two-person household model is a natural setting in which to also consider home production, we will focus on the basic model without home production in order to minimize notation.

For ease of exposition we will refer to our two members as male and female, using subscripts m and f respectively. The household allocation will consist of total consumption and hours of work for each member. Once again we assume a separable utility function:

$$U(c, h_f, h_m) = u(c) - v(h_f) - v(h_m),$$

where for simplicity we have assumed that disutility from working takes the same form for both members. The household now faces the following budget equation:

$$c = w_f h_f + w_m h_m + I,$$

where w_j is the wage rate that member j faces, and as before, I is non-labor income of the household. Another effect is present in this model beyond the income and substitution effects that we have previously noted. Specifically, the labor supply of each member depends

[4]See Bourguignon and Chiappori (1992) for an early discussion and analysis of models that explicitly model households as composed of distinct individuals.

not only upon non-labor income and their own wage, but also on the wage of the other household member. To see this, we derive the first order conditions for the two choices of labor supply:

$$\begin{aligned} h_f &: u'(w_f h_f + w_m h_m + I) w_f = v'(h_f) \\ h_m &: u'(w_f h_f + w_m h_m + I) w_m = v'(h_m). \end{aligned}$$

Looking at the first equation, it is apparent that higher labor earnings of the male member act like an increase in non-labor income and thus will lead to lower labor supply of the female member. And similarly, higher labor earnings for the female member will decrease the labor supply of the male member. This indicates that the two labor supplies are jointly determined. Dividing the two FOCs by each other gives:

$$\frac{v'(h_f)}{v'(h_m)} = \frac{w_f}{w_m}.$$

This implies that relative wages are positively correlated with relative hours. Our functional form for v implies the following relationship between relative hours and relative wages:

$$\frac{h_f}{h_m} = \left[\frac{w_f}{w_m}\right]^{\gamma}.$$

Previously we showed how the value of the curvature parameter γ affected the response of labor supply of a single individual to changes in their own wage rate. In this setting this parameter also dictates the extent to which the household reallocates work across household members in response to changes in relative wages. An immediate implication is that the model has a force that can lead to specialization within the household, with the higher wage individual engaging in much more market work. If we were to generalize this setting to explicitly include home production as an activity, we would obtain the prediction that holding all else constant, differences in market wages are a force that leads to specialization in home and market work within the household. This force predicts that an exogenous reduction in the gender wage gap will lead to an increase in the labor supply of married females relative to married males.[5]

This extension also provides insight into a phenomenon known as the added worker effect. This describes the tendency for a secondary earner within a family to increase their labor supply when the primary earner experiences an unemployment spell. While our labor supply model does not account for unemployment, it does explain why an exogenous reduction in h_m will imply an increase in h_w.

The productivity of leisure activities

In our discussion thus far, we have specified preferences in terms of the disutility of work, so that leisure has been somewhat in the background. In this subsection we consider an alternative specification that places more emphasis on leisure. Previously we defined utility over consumption and hours and wrote:

$$U(c, h) = u(c) - v(h).$$

[5]See, for example the analysis in Attanasio, Low, and Sánchez-Marcos (2008).

In this section we will instead define utility over consumption and leisure. Leisure will be defined as the difference between total discretionary time and time spent working. Denoting leisure by l, and letting $\bar{h}$ be the total amount of discretionary time, we have that:

$$l = \bar{h} - h$$

where as before, h is time devoted to market work. Again assuming preferences that are separable, we will now write utility as:

$$\tilde{U}(c, l) = u(c) + \tilde{v}(l),$$

where $\tilde{v}$ is now a strictly concave function. Analogous to our earlier analysis, a commonly used functional form in this case is:

$$\tilde{v}(l) = \frac{1}{1 - 1/\gamma} l^{1-1/\gamma},$$

where $\gamma > 0$.

While the two specifications are very similar, there are some differences worth noting. First, in the previous specification, we noted that the solution for h was always interior, independently of the value for non-labor income I. This followed from the fact that the marginal disutility of work was equal to zero when $h = 0$. This is no longer the case. For example, if non-labor income is sufficiently high, it is possible that $h = 0$ is optimal. Second, the previous specification implied that the elasticity of utility with respect to work, i.e., $hv'(h)/v(h)$ was constant. This is no longer true, as the elasticity of the utility with respect to time spent working is now given by $-h\tilde{v}'(\bar{h} - h)/(\bar{h} - h)v(\bar{h} - h)$. This value approaches infinity as h approaches $\bar{h}$ and leisure tends to zero.

Starting from the specification in which we write utility as $\tilde{U}$, we now consider a simple extension to illustrate the point made by Aguiar, Bils, Hurst, and Charles (2021) regarding how changes in technology associated with leisure activities will directly impact labor supply choices. To do this we dig deeper into the specification of how leisure produces utility. In particular, we assume that there are J different leisure activities, which we index by j. An individual allocates their total leisure time l across the J different leisure activities, and we denote by l_j the amount of leisure time denoted to leisure activity j. In the simplest specification, total utility from leisure is then written as:

$$\tilde{v}\left(\sum_j \frac{(A_j l_j)^{1-\delta}}{1-\delta}\right),$$

where the A_j can be interpreted as the productivity of leisure time in leisure activity j. In this setting, increases in some or all of the A_j will mechanically act like an increase in the value of ψ in our original specification of utility. At this level, one might think that this is purely a game of words about whether we label the change in ψ a change in preferences versus a change in the productivity of leisure. The key contribution of Aguiar et al. (2021) was to develop a structure that allows us to examine this more deeply. A full analysis is beyond the scope of what we will cover here, but we can provide a few details. By examining the evolution of trend changes in the allocation of leisure relative to how the allocation of

leisure time changes during a large downturn in economic activity, the authors were able to assess the extent to which the trend changes reflect changes in the A_j. One of their key findings is that there was a large increase in the productivity of leisure time allocated to computer and video games, and that this in turn led to a large increase in the total amount of time devoted to leisure time by young males.

12.3.3 Dynamic labor supply: a first look

Our analysis of the basic static labor supply problem emphasized wage and income elasticities as two important forces that shape the response of labor supply to changes in the economic environment. In dynamic settings there is another elasticity that is important for understanding labor supply responses. This will feature prominently in later parts of this chapter, but we first introduce it here by extending our previous analysis to a two period setting.

We now assume that an individual has preferences over consumption and labor supply given by:

$$U(c_1, c_2, h_1, h_2) = \sum_{t=1}^{2} \beta^{t-1} \left[\frac{1}{1-\sigma} c_t^{1-\sigma} - \frac{\psi}{1+1/\gamma} h_t^{1+1/\gamma} \right],$$

where the parameters σ, γ, and ψ are as before and $\beta \in (0,1)$ is the discount factor. The individual faces a real wage of w_1 in the first period and w_2 in the second period. Importantly, we also assume that the individual can borrow and/or save at the interest rate r, and thus faces a lifetime budget constraint given by:

$$c_1 + \frac{c_2}{1+r} = w_1 h_1 + \frac{w_2 h_2}{1+r} + I,$$

where I is now initial non-labor wealth.[6]

If $\beta = 1/(1+r)$ and $w_1 = w_2$ the optimal choice of h is the same in both periods and is essentially the same as analyzed in the previous subsection, modulo normalizing non-labor income appropriately. Our interest in this subsection is in the situation in which there are differences in wages in the two periods. To pursue this we derive first order conditions for the individual lifetime maximization problem. Letting λ be the Lagrange multiplier on the lifetime budget constraint, and focusing on the case in which $\beta = 1/(1+r)$ we obtain the following first order conditions:

$$\begin{aligned} c_1 &: c_1^{1-\sigma} = \lambda \\ c_2 &: c_2^{1-\sigma} = \lambda \\ h_1 &: \psi h_1^{1/\gamma} = \lambda w_1 \\ h_2 &: \psi h_2^{1/\gamma} = \lambda w_2. \end{aligned}$$

Note that the value of the Lagrange multiplier λ is the marginal utility of consumption. Dividing the last two equations by each other one obtains:

$$\left[\frac{h_1}{h_2} \right]^{1/\gamma} = \frac{w_1}{w_2},$$

[6] If the individual is not able to move purchasing power across time by either saving or borrowing then the individual will effectively solve two static problems of the form that we previously studied.

or taking logs,

$$\log h_1 - \log h_2 = \gamma \left[\log w_1 - \log w_2\right].$$

Recalling that $\gamma > 0$, this expression indicates that hours and wage rates are always positively correlated intertemporally. Loosely speaking, because purchasing power can be transferred across time, the individual should work relatively more when wages are relatively high. The strength of this effect is dictated solely by the preference parameter γ, and for this reason γ is referred to as the intertemporal elasticity of substitution for labor supply. This elasticity, which holds the marginal utility of consumption fixed, is also known as the **Frisch elasticity**. This elasticity plays a key role in understanding the role of labor supply in the context of business cycle fluctuations. This equation has played a key role in work that estimates the Frisch elasticity using panel data on individuals.[7]

12.4 Dynamic models of labor supply: theory and estimation

We now present a fully dynamic stochastic life-cycle model, and generalize our derivations for the three labor supply elasticities (Marshallian, Hicksian, and Frisch) in Section 12.3. The key challenge, relative to the static model, is that non-labor income is now endogenous because it contains savings, a choice variable for the household. To overcome this challenge, we use the so-called "two-stage budgeting" approach (see Blundell and MaCurdy, 1999 and Keane, 2011). This method splits the dynamic optimization problem of the household in two stages. In the first stage, the household allocates its lifetime resources intertemporally, by choosing savings in each period. Once savings are chosen, non-labor income is given, and the second-stage problem, where the household chooses hours worked, is analogous to the static model

Consider an individual i who lives for T periods (where $T = \infty$ is a special case), has discount factor $\beta > 0$, and derives utility from consumption c_{it} and disutility from hours worked h_{it}. We normalize the time endowment to 1. Let w_{it} be the hourly wage of the individual at date t. Wages fluctuate stochastically and are a source of uncertainty for individuals. The individual can trade a risk-free asset a_{it} with constant gross rate of return $R = 1 + r$, subject to a borrowing limit $\underline{a}_i$. The government levies a proportional tax τ on labor income and pays a lump-sum transfer $\mathcal{T}$.

For a given initial wealth a_{i0}, the optimization problem of the individual is

$$\max_{\{c_{it}, h_{it}\}_{t=0}^{T}} \mathbb{E}_0 \sum_{t=0}^{T} \beta^t U\left(c_{it}, h_{it}\right) \tag{12.7}$$

subject to

$$c_{it} + a_{i,t+1} = R a_{it} + (1 - \tau) w_{it} h_{it} + \mathcal{T},$$
$$a_{i,t+1} \geq -\underline{a}_i, \quad c_{it} \geq 0, \text{ and } 0 \leq h_{it} \leq 1.$$

[7] For future reference we note that combining the equation for consumption and labor in period t one obtains the equation $\log h_t = (\gamma/\psi) + \gamma \log w_t + \gamma(1-\sigma) \log c_t$, which can also be used to provide an estimate of γ if one has data on hours, wages and consumption.

In addition, we require the No-Ponzi condition $a_{i,T+1} \geq 0$ which, in the infinite horizon version of the model, becomes $\lim_{T\to\infty} a_{i,T}/R^T \geq 0$, as explained in Section 4.3.1.

For much of this section, it is useful to specialize to the same preferences used in Section 12.3:

$$U(c_{it}, h_{it}) = \frac{c_{it}^{1-\sigma} - 1}{1-\sigma} - \psi \frac{h_{it}^{1+1/\gamma}}{1+1/\gamma} \tag{12.8}$$

with $\sigma \geq 0$ and $\gamma \geq 0$. Recall that $1/\sigma$ is the intertemporal elasticity of substitution for consumption.

12.4.1 Derivations of labor supply elasticities

We start by defining non-labor income net of saving as

$$I_{it} = ra_{it} + \mathcal{T} - (a_{i,t+1} - a_{it}). \tag{12.9}$$

This definition is convenient because it allows us to rewrite the time t budget constraint as

$$c_{it} = (1-\tau) w_{it} h_{it} + I_{it}, \tag{12.10}$$

which is isomorphic to the formulation in the static model of Section 12.3. This is the cornerstone of the two-state budgeting method. To fully understand this approach, normalize the time endowment to 1, and rewrite the budget constraint as

$$c_{it} + (1-\tau) w_{it} (1 - h_{it}) = Y_{it}$$

where Y_{it} is "full income", i.e. the maximum possible income earned by the individual at time t taking $a_{i,t+1}$ as given, or

$$Y_{it} = (1-\tau) w_{it} + R_t a_{it} + \mathcal{T} - a_{i,t+1}. \tag{12.11}$$

The two-state budgeting approach exploits the idea that one can split the dynamic optimization problem of the household (12.7) in two steps. First, the household decides how to optimally allocate its life-time full income period by period. This amounts to choosing a sequence $\{a_{i,t+1}\}_{t=0}^{T}$ based on the Euler equation, the intertemporal optimality condition that describes the trade-off between consuming today and consuming next period.[8] Next, given $a_{i,t+1}$, the individual chooses (c_{it}, h_{it}) at every t based on their intratemporal first-order condition and the budget constraint. Thus, in what follows, we focus on a generic time t and take $a_{i,t+1}$ as given.

Marshallian Elasticity. We start by deriving the expression for the Marshallian elasticity $\varepsilon_{h,w}^{M}$. Recall from the discussion of Section 12.3 that this is the total elasticity of hours worked h_{it} to a change in the wage w_{it}. By total, we mean that it incorporates both the substitution and the income effect. Namely, because in calculating it we do not compensate the change in the wage with a corresponding change in non-labor income of the opposite sign, i.e. we keep I_{it} fixed, this elasticity is also called "uncompensated".

[8] We derived and discussed the Euler equation for consumption in Chapter 4.

Start by differentiating the budget constraint (12.10) with respect to w_{it}, c_{it} and h_{it}. Rearranging terms, we obtain:

$$\frac{d\log c_{it}}{d\log w_{it}} = \left(1 - s_{it}^{I}\right) + \left(1 - s_{it}^{I}\right)\frac{d\log h_{it}}{d\log w_{it}}, \tag{12.12}$$

where $s_{it}^{I} = I_{it}/c_{it} < 1$ is the non-labor income share of consumption expenditures. The intratemporal first order condition with respect to h_{it} yields

$$c_{it}^{-\sigma} w_{it}\left(1-\tau\right) = \psi h_{it}^{\frac{1}{\gamma}}. \tag{12.13}$$

Differentiating (12.13) with respect to w_{it}, c_{it} and h_{it}, we arrive at:

$$\frac{1}{\sigma}\frac{d\log c_{it}}{d\log w_{it}} + \frac{1}{\gamma}\frac{d\log h_{it}}{d\log w_{it}} = 1. \tag{12.14}$$

Combining (12.12) and (12.14), we obtain the final expression for the Marshallian elasticity

$$\varepsilon_{h,w}^{M} \equiv \frac{d\log h_{it}}{d\log w_{it}} = \frac{1 - \left(1 - s_{it}^{I}\right)/\sigma}{1/\gamma + \left(1 - s_{it}^{I}\right)/\sigma} \tag{12.15}$$

which, modulo the different definition of s_{it}^{I} adapted to the dynamic model, is the same as equation (12.3) in the static model of Section 12.3. Once again, note that $\varepsilon_{h,w}^{M}$ can be positive or negative, depending on whether the substitution effect is larger or smaller than the income effect.

Hicksian elasticity. We now turn to the Hicksian or "compensated" elasticity. This elasticity measures how hours worked optimally respond to a wage change that is compensated by an equal change in non-labor income of the opposite sign. This compensation offsets the income effect from the wage change, thus the Hicksian elasticity only captures the substitution effect that leads individuals to allocate more time to work (less to leisure) when the wage (the price of leisure) goes up.

Let $h_{it}\left(w_{it}, I_{it}\right)$ denote optimal hours worked expressed as a function of wage and net non-labor income from the first order condition (12.13). Differentiating this function, we obtain:

$$dh_{it}\left(w_{it}, I_{it}\right) = \frac{\partial h_{it}\left(w_{it}, I_{it}\right)}{\partial w_{it}} dw_{it} + \frac{\partial h_{it}\left(w_{it}, I_{it}\right)}{\partial I_{it}} dI_{it},$$

and rearranging, we arrive at:

$$\frac{dh_{it}\left(w_{it}, I_{it}\right)}{dw_{it}}\frac{w_{it}}{h_{it}} = \frac{\partial h_{it}\left(w_{it}, I_{it}\right)}{\partial w_{it}}\frac{w_{it}}{h_{it}} + \frac{\partial h_{it}\left(w_{it}, I_{it}\right)}{\partial I_{it}}\frac{I_{it}}{h_{it}}\left(\frac{dI_{it}}{dw_{it}}\frac{w_{it}}{I_{it}}\right).$$

The term on the left-hand side is the Marshallian elasticity. The first term on the right-hand side can be interpreted as the Hicksian elasticity (i.e., the pure substitution effect) as long as the second term exactly compensates the individual with a change in non-labor income equal to the labor income shift, or as long as $dI_{it}/dw_{it} = \left(1-\tau\right)h_{it}$. Thus, expressing the equation above in terms of elasticities, we arrive at the so-called Slutsky equation:

$$\varepsilon_{h,w}^{H} = \varepsilon_{h,w}^{M} - \varepsilon_{h,I}\left(\frac{1 - s_{it}^{I}}{s_{it}^{I}}\right), \tag{12.16}$$

which is the dynamic counterpart of equation (12.4) in the static model of Section 12.3.

To obtain the elasticity of hours worked to net non-labor income, differentiate (12.13) with respect to h_{it} and I_{it}. Simple algebra yields

$$\varepsilon_{h,I} \equiv \frac{d \log h_{it}}{d \log I_{it}} = \frac{-s_{it}/\sigma}{1/\gamma + (1 - s_{it})/\sigma}$$

which, substituted into (12.16), delivers the Hicksian labor supply elasticity

$$\varepsilon^H_{h,w} = \frac{1}{1/\gamma + (1 - s^I_{it})/\sigma}. \tag{12.17}$$

Once again, note that the Hicksian is always positive. In addition, $\varepsilon^H_{h,w} \geq \varepsilon^M_{h,w}$ where equality holds in the absence of income effects, i.e. when period utility is linear in consumption, or $\sigma \to \infty$.

The fact that the expressions for Marshallian and Hicksian elasticities are the same in the static and dynamic model suggests that these are fundamentally static concepts, as both hold constant the intertemporal allocation of resources. These concepts are, therefore, especially useful to analyze shifts in wages or in taxes that are perceived as being permanent by households. Examples of the former are a permanent, or very persistent, increase in individual labor productivity, a promotion associated with a pay raise, or a move to a better-paid job. Examples of the latter are income tax reforms that are expected to last indefinitely. In the absence of binding borrowing constraints, following this sort of wage increases (or a tax cuts), the household would spend roughly all the extra income every period and there would be no, or little, change in saving patterns.[9] If, however, we want to analyze temporary wage changes (e.g., a seasonal bonus) or income tax changes (e.g., tax reforms with expiring provisions) that lead to a shift in saving patterns, we need a different concept of elasticity. This motivates us to introduce the concept of the Frisch elasticity.

Frisch elasticity. The Frisch elasticity of labor supply describes the response of hours worked to a change in the wage keeping the marginal utility of wealth constant. It is the relevant concept to analyze how either a transitory or an anticipated wage change affects hours. Small enough temporary wage fluctuations have only a negligible effect on lifetime wealth, and anticipated wage changes convey no new information. Thus neither one impacts the marginal utility of wealth. As a result, the Frisch elasticity is especially useful to assess the implications of wage fluctuations over the business cycle or of temporary tax changes.

It is easy to derive an expression for the Frisch elasticity for generic period utility $U(c_{it}, h_{it})$ without specializing to a particular functional form for now. Let the marginal utility of wealth for individual i at date t be λ_{it}. The first-order conditions with respect to consumption and hours worked for the problem specified previously in (12.7) are (U_c is the marginal utility of consumption, U_h is the marginal disutility of working, and U_{ij}, $i, j = c, h$ represents a second derivative)

$$U_c = \lambda_{it} \tag{12.18}$$

[9] There might be, however, a change in non-labor income if, for example, some of the additional tax reveues are redistributed as lump-sum transfers to households. As in the static model, this makes the Hicksian the relevant elasticity.

and

$$-U_h = \lambda_{it}(1-\tau)w_{it}. \tag{12.19}$$

Differentiating (12.19), while keeping λ_{it} constant, yields

$$-h_{it}U_{hh}\frac{dh_{it}}{h_{it}} - U_{hc}dc_{it} = \lambda_{it}(1-\tau)w_{it}\frac{dw_{it}}{w_{it}}.$$

Using (12.19) again to substitute out $\lambda_{it}(1-\tau)w_{it}$ and rearranging leads to:

$$h_{it}U_{hh}\frac{dh_{it}}{h_{it}} + h_{it}U_{hc}\frac{dc_{it}}{dh_{it}}\frac{dh_{it}}{h_{it}} = U_h\frac{dw_{it}}{w_{it}}. \tag{12.20}$$

Differentiating (12.18) gives

$$\frac{dc_{it}}{dh_{it}} = \frac{-U_{ch}}{U_{cc}}$$

which, substituted into (12.20), gives the general expression for the Frisch elasticity:

$$\varepsilon^F_{h,w} = \frac{U_h}{h_{it}U_{hh} - h_{it}(U^2_{hc}/U_{cc})}. \tag{12.21}$$

It is useful to derive the Frisch elasticity for some particular functional forms of the period utility function. For (12.8), it is easy to see from (12.21) that $\varepsilon^F_{h,w} = \gamma$. It is now clear why (12.8) is such a common preference specification in macroeconomics: its two curvature parameters fully control two key elasticities, the intertemporal elasticity of substitution $(1/\sigma)$ and the Frisch elasticity (γ). Comparing (12.15), (12.17) and (12.21) for the separable utility (12.8) we conclude that

$$\varepsilon^F_{h,w} \geq \varepsilon^H_{h,w} \geq \varepsilon^M_{h,w}$$

with equality holding when $\sigma \to \infty$. In this case, utility becomes quasi-linear and the income effect vanishes.

Consider now the **GHH utility function** introduced by Greenwood, Hercowitz, and Huffman (1988)

$$U(c_{it}, h_{it}) = \frac{\left(c_{it} - \psi\frac{h_{it}^{1+1/\gamma}}{1+1/\gamma}\right)^{1-\sigma} - 1}{1-\sigma}. \tag{12.22}$$

Applying (12.21) yields again that $\varepsilon^F_{h,w} = \gamma$. Another key property of these preferences is that the income effect is zero. Thus, for GHH utility, $\varepsilon^M_{h,w} = \varepsilon^H_{h,w} = \varepsilon^F_{h,w} = \gamma$.

We conclude by noting that not all utility functions imply a constant Frisch elasticity. For example, for the specifications

$$U(c_{it}, h_{it}) = \frac{\left(c_{it}^{1-\psi}(1-h_{it})^{\psi}\right)^{1-\sigma} - 1}{1-\sigma} \quad \text{and} \quad U(c_{it}, h_{it}) = \frac{c_{it}^{1-\sigma} - 1}{1-\sigma} + \psi\frac{(1-h_{it})^{1-\gamma}}{1-\gamma}$$

it is the Frisch elasticity of leisure that is constant, while the Frisch elasticity of hours now depends on hours worked h_{it}. For the first specification, we have

$$\varepsilon^F_{h,w} = \left(\frac{1-h_{it}}{h_{it}}\right)\left(\frac{1-(1-\psi)(1-\sigma)}{\sigma}\right)$$

and for the second

$$\varepsilon^F_{h,w} = \left(\frac{1-h_{it}}{h_{it}}\right)\frac{1}{\gamma}.$$

Thus, in both cases, individuals who work more hours are less responsive to temporary wage changes.

Expression (12.21) was derived only under the assumption that individuals optimize intratemporally. Nothing was explicitly assumed about intertemporal saving behavior, nor about asset market structure. One could be, therefore, tempted to conclude that, as long as equation (12.19) holds, observing the response of hours worked to a transitory wage change identifies γ correctly. This deduction is, however, incorrect. For example, for a constrained individual even a temporary increase in labor income affects the marginal utility of wealth and that, in turn, affects optimal labor supply.

To further appreciate this point, and illustrate the intertemporal nature of this elasticity, note that abstracting from uncertainty and in the absence of borrowing constraints, under separable utility (12.8) , the Euler equation for problem (12.7) yields

$$c_{it}^{-\sigma} = \beta R c_{i,t+1}^{-\sigma}. \tag{12.23}$$

Substituting the intratemporal first-order condition (12.13) into (12.23) we obtain

$$\frac{d\log\left(h_{i,t+1}/h_{it}\right)}{d\log\left(w_{i,t+1}/w_{it}\right)} = \gamma.$$

Thus, as shown for the 2-period model in Section 12.3, the Frisch elasticity dictates the response of relative hours worked across two periods with respect to changes in relative wages. If taxes were time varying, γ would also describe the elasticity of relative hours to relative changes in taxes between the two periods. The larger is γ the more individuals will shift hours worked in response to temporary wage or tax changes. Importantly, this derivation hinges on the fact that equation (12.23) holds with equality, i.e. liquidity constraints do not bind, and there is no uncertainty and thus no precautionary motive. In the next section, we discuss how the presence of these two factors complicate the estimation of γ.

12.4.2 Estimation of Frisch elasticity from micro data

In the separable utility specification, commonly used in macroeconomics, the two key elasticity parameters are σ and γ, so historically there has been great interest in estimating both of them. The parameter σ governs both the intertemporal elasticity of substitution for consumption and risk aversion. Under the former interpretation, $1/\sigma$ measures the sensitivity of consumption growth to the interest rate. Under the second one, it measures, for example, how households allocate their portfolio between risky and safe assets, for a given equity premium, or how large the equity premium can be, for a given volatility of aggregate consumption (see Chapter 7). Thus, the estimation of σ has been mostly the domain of the consumption and asset pricing literatures.[10] For this reason, here we focus on γ and present

[10] As we discuss later in this chapter, conditional on assuming that utility is separable between consumption and hours, macroeconomists have often focused on the case in which $\sigma = 1$ since it is the only value consistent with a balanced growth path solution that features constant hours.

a brief overview of the empirical challenges underlying the estimation of the Frisch elasticity from micro data, and how the macro-labor literature has evolved over time to tackle these challenges.

In what follows, we assume that households solve (12.7) and have rational expectations, i.e. they know the stochastic process underlying fluctuations in w_{it}. In addition, we assume that utility takes the separable functional form in (12.8), therefore the parameter to be estimated is γ. We discuss two complementary approaches to the measurement of γ, a structural approach and an experimental one.

Structural approach

Taking logs of the two optimality conditions (12.18) and (12.19), and using the notation $const_i$ for individual specific constant terms, we obtain:

$$\log c_{it} = -\frac{1}{\sigma} \log \lambda_{it} \tag{12.24}$$

and

$$\log h_{it} = const_i + \gamma \log w_{it} + \gamma \log \lambda_{it}. \tag{12.25}$$

The minimum requirement to obtain an empirical estimate of γ from equation (12.25) is the availability of longitudinal micro data on individual hourly wages and hours worked. The key challenge is that λ_{it}, the marginal utility of consumption for individual i, is unobservable. Suppose we naively try to estimate equation (12.25) by OLS, possibly using individual fixed effects to absorb the individual-specific constant term. This approach would be problematic because λ_{it}, which enters the residual, is likely to be correlated with the wage when the latter has a persistent component that affects wealth and consumption. In this case, $cov\left(\log w_{it}, \log \lambda_{it}\right) < 0$, which induces a downward bias in the estimator $\hat{\gamma}$. The same logic applies if the estimation is done in first differences. The empirical literature has tackled this problem in three ways, by using (i) instrumental variables, (ii) data on expected wage growth, and (iii) data on consumption expenditures.

Instrumental variables. The most natural way to address this potential bias in OLS is through instrumental variables. Express (12.25) in first differences between time $t-1$ and time t

$$\Delta \log h_{it} = \gamma \Delta \log w_{it} + \gamma \Delta \log \lambda_{it}. \tag{12.26}$$

To substitute out the Lagrange multiplier note that, from household optimization, the optimal saving decision yields the following Euler equation expressed in logarithms

$$\log\left(\lambda_{it} - \phi_{it}\right) = const_i + \log \mathbb{E}_t\left[\lambda_{i,t+1}\right], \tag{12.27}$$

where ϕ_{it} is the the multiplier on the borrowing constraint. Assuming that λ_{it} is conditionally log-normally distributed, then

$$\log \mathbb{E}_t\left[\lambda_{i,t+1}\right] = \mathbb{E}_t\left[\log \lambda_{i,t+1}\right] + \frac{1}{2} \mathrm{Var}_t\left(\log \lambda_{i,t+1}\right), \tag{12.28}$$

where Var_t is the conditional variance. Under rational expectations, we can substitute (12.28) into (12.27) and write (12.27) as

$$\log(\lambda_{it} - \phi_{it}) = const_i + \log \lambda_{i,t+1} - \xi_{i,t+1} + \frac{1}{2}\text{Var}_t(\log \lambda_{i,t+1}),$$

where $\xi_{i,t+1}$ is the prediction error for the marginal utility of consumption which is orthogonal to all information available to worker i at time t. Using a first-order approximation of the left-hand-side around $\phi^* = 0$, and recognizing that $\text{Var}_t(\log \lambda_{i,t+1}) = \text{Var}_t(\xi_{i,t+1})$, we obtain

$$\log \lambda_{it} - \frac{\phi_{it}}{\lambda_{it}} = const_i + \log \lambda_{i,t+1} - \xi_{i,t+1} + \frac{1}{2}\text{Var}_t(\xi_{i,t+1}).$$

This equation can be used to substitute out the Lagrange multiplier λ_{it} in (12.26) and obtain

$$\Delta \log h_{it} = \gamma \Delta \log w_{it} + \gamma \left[\xi_{it} - \frac{\phi_{i,t-1}}{\lambda_{i,t-1}} - \frac{1}{2}\text{Var}_{t-1}(\xi_{it})\right]. \tag{12.29}$$

Once again, one would expect the residual ξ_{it} to be negatively correlated with $\Delta \log w_{it}$: positive news about wage growth, unless very transitory, would translate into an unexpected growth in consumption, and hence a decline in ξ_{it}.

MaCurdy (1981) abstracts from potentially binding liquidity constraints and uninsurable uncertainty, i.e. assumes $\phi_{i,t-1} = 0$ and $\text{Var}_{t-1}(\xi_{it}) = 0$, and uses lagged wage growth $\Delta \log w_{i,t-1}$ as an instrument to deal with the correlation between ξ_{it} and $\Delta \log w_{it}$ in equation (12.29). In practice, this is a relatively ineffective strategy for two reasons. First, in surveys hourly wages are typically measured as earnings divided by hours worked, but hours are notoriously measured with error. By itself, this yields a downward bias in the estimate of γ. This is the first reason why lagged wage growth is a weak instrument. The second one is that if the statistical process for wages is very persistent, then the autocorrelation of wage growth with past wage growth is very small –for example, in the limiting case in which wages follow a random walk this autocorrelation is exactly zero.

Wage growth expectations. One can always write

$$\Delta \log w_{it} = \mathbb{E}_{t-1}[\Delta \log w_{it}] + \varepsilon_{it} \tag{12.30}$$

where, under rational expectations, the shock ε_{it} is orthogonal to predicted wage growth $\mathbb{E}_{t-1}[\Delta \log w_{it}]$. Substituting (12.30) into (12.26), we obtain

$$\Delta \log h_{it} = \gamma \mathbb{E}_{t-1}[\Delta \log w_{it}] + \gamma(\varepsilon_{it} + \Delta \log \lambda_{it}). \tag{12.31}$$

Under the permanent-income hypothesis –i.e., once again abstracting from liquidity constraints– consumption growth between $t-1$ and t, and therefore $\Delta \log \lambda_{it}$, is only affected by the news accruing over that period, i.e. by ε_{it}, but not by anything which is already incorporated in the information set at time $t-1$. As a result, with an external measure of expected wage growth one could estimate γ consistently. Pistaferri (2003) estimates equation (12.31) directly using Italian survey data which contain information about individual subjective earnings expectations, and obtains an estimate of $\gamma = 0.7$.

Domeij and Floden (2006) point out that both the wage growth expectations approach and the IV strategy fail if there are binding borrowing constraints. Consider first equation (12.31). Individuals with very steep predictable wage growth are more likely to be constrained because they would like to consume more out of their future income stream, which induces a negative correlation between $\mathbb{E}_{t-1}[\Delta \log w_{it}]$ and the growth in the marginal utility of consumption $\Delta \log \lambda_{it}$. This argument keeps holding in equation (12.29) if one uses $\Delta \log w_{i,t-1}$ as an instrument for $\Delta \log w_{it}$. In addition, there is another form of bias coming from constrained agents. In the absence of liquidity constraints, households choose to work hard in periods with temporarily high wages and to enjoy more leisure when wages are low. However, when borrowing constrained agents receive transitory negative income shocks, in order to avoid cutting consumption drastically they will increase labor supply in response to falling wages. This force is another source of downward bias in the estimation of the Frisch elasticity.

Low (2005) advances a related critique. Even in the absence of binding constraints, uninsurable income risk leads to a precautionary saving motive, as shown in Chapter 11. One way in which workers can increase their precautionary saving is to work more and use the additional earnings to increase saving, a sort of precautionary labor supply motive. From equation (12.29) it is clear that if this motive is strong (i.e., $\text{Var}_{t-1}(\xi_{it})$ is large) when wage growth is high, precautionary labor supply is yet another source of downward omitted variable bias. This would be the case, for example, if young workers who typically face steep wage growth also hold low amounts of wealth.

Consumption data. If consumption expenditures are directly observable in the data, one can exploit equation (12.24) to use consumption in place of λ_{it}. Combining (12.24) and (12.26) and expressing the resulting equation in first differences yields

$$\Delta \log h_{it} = \gamma \Delta \log w_{it} - \gamma \sigma \Delta \log c_{it}.$$

Altonji (1986b) implements this approach for the U.S. using data on the Panel Study of Income Dynamics (PSID), a panel dataset with information on earnings, hours worked, and expenditures on some selected consumption goods (e.g., food, clothing, and shelter). The advantage of this approach is that it is based on the envelope condition which always holds with equality, so it is immune to binding borrowing constraint. Its key limitation is that consumption expenditures in surveys are measured with error and, as a result, can have a weak correlation with the true marginal utility of consumption. A second problem is that food is often used as a proxy for total consumption because it is better measured and more commonly present in surveys. However, it only represents a relatively small share of the budget for most households.

Experimental and quasi-experimental evidence

In light of the challenges in estimating the Frisch elasticity through structural approaches, the literature has explored, in parallel, RCTs or quasi-exogenous empirical variation that can shed light on this key parameter. One example of the former approach is Fehr and Goette (2007). They conducted a randomized field experiment at a bicycle messenger service in Zurich, Switzerland. The bicycle messengers are paid solely on commission, i.e. they retain

a share of the revenues they generate. The authors implemented an exogenous and transitory increase of 25 percent in the rate of commission for a random subgroup of workers. Since the wage was increased only during four weeks, its impact on the workers' lifetime wealth is negligible. By comparing the treated group with the control group, the authors estimated an intertemporal elasticity of labor supply which exceeds 1. Bianchi, Gudmundsson, and Zoega (2001) is an example of quasi-exogenous variation from the macro data. This paper exploits a tax-reform in Iceland resulting in a year, 1987, free of labor income taxes. This tax-free year created a strong incentive for intertemporal substitution of work, but a minimal income effect because the reform applied only to one specific year, and thus did only imply a small change in life-time resources. As a result, the tax-free year offers a rare natural experiment suitable to estimate the Frisch elasticity. The authors estimate an average Frisch elasticity around 0.4.

Optimization frictions

Optimization frictions occur when households encounter costs associated with updating previously optimal decisions in light of the new environment. Examples are physical costs of adjusting the choice variable, costs of acquiring new information, or behavioral biases such as the so called "status-quo" bias. Chetty (2012) pointed out that these frictions can pose a challenge to estimate labor supply elasticities from microdata.

To illustrate this point, we consider a static model of a household with GHH preferences. This preference specification, albeit non-standard, allows us to abstract from income effects— which simplifies the derivations— and to focus on the role of optimization frictions. Consider the static optimization problem of a household:

$$U\left(c^*, h^*\right) = \max_{c,h} \log\left(c - \psi \frac{h^{1+\frac{1}{\gamma}}}{1+\frac{1}{\gamma}}\right) - \Delta \mathbb{I}_{\{h \neq h^*\}} \tag{12.32}$$

subject to

$$c = wh.$$

Recall that under these preferences Frisch, Marshallian, and Hicksian elasticities are all the same and equal to γ. Note the fixed re-optimization cost $\Delta > 0$, expressed in utility terms, which kicks in whenever the choice of hours worked differs from its current level which we assume to be at the optimum, given the current wage w. It is easy to see that the optimal choices of hours worked h^* and consumption c^* when the wage equals w are:

$$h^* = \left(\frac{w}{\psi}\right)^{\gamma} \tag{12.33}$$

and

$$c^* = wh^*,$$

and utility evaluated at the optimum is

$$U\left(c^*, h^*\right) = -\log\left(1+\gamma\right) - \gamma \log \psi + \left(1+\gamma\right) \log w,$$

which is increasing in the wage and decreasing in the disutility of labor ψ, as expected.

Consider now the effect of an increase in the wage from w to $w(1+\varepsilon)$ on utility and, for the moment, ignore re-optimization costs, i.e. assume the usual frictionless optimization environment. This utility gain can be decomposed into the direct effect of the wage change, holding h fixed at the old choice h^*, plus the effect induced by the behavioral response of reoptimizing labor supply, keeping consumption fixed at the new choice:

$$U(c_\varepsilon^*, h_\varepsilon^*) - U(c^*, h^*) = [U(c_\varepsilon^*, h^*) - U(c^*, h^*)] + [U(c_\varepsilon^*, h_\varepsilon^*) - U(c_\varepsilon^*, h^*)],$$

where the notation $(c_\varepsilon^*, h_\varepsilon^*)$ denotes the updated choices after the wage increase, and (c^*, h^*) denotes the old choice before the wage changed.

From (12.32) evaluated at (c_ε^*, h^*) and (c^*, h^*) one obtains that

$$U(c_\varepsilon^*, h^*) - U(c^*, h^*) = \log(1 + \varepsilon(1+\gamma)),$$

and note that this gain is not costly to achieve, so households will always update their consumption decision, conditional on their choice of hours worked. From (12.32) evaluated at $(c_\varepsilon^*, h_\varepsilon^*)$ and (c_ε^*, h^*) one obtains that

$$U(c_\varepsilon^*, h_\varepsilon^*) - U(c_\varepsilon^*, h^*) = (1+\gamma)\log(1+\varepsilon) - \log(1+\varepsilon(1+\gamma)). \tag{12.34}$$

This is the term we are interested in because it represents the utility gain from re-optimizing the choice of hours, which is costly. This term is different from zero only up to second order. To see this, consider a second order Taylor expansion of both terms on the right-hand side of (12.34) around $\varepsilon = 0$ (or $h_\varepsilon^* = h^*$):

$$\begin{aligned}(1+\gamma)\log(1+\varepsilon) &\simeq (1+\gamma)\varepsilon - \frac{1}{2}(1+\gamma)\varepsilon^2 \\ \log(1+\varepsilon(1+\gamma)) &\simeq (1+\gamma)\varepsilon - \frac{1}{2}(1+\gamma)^2\varepsilon^2\end{aligned}$$

and note that up to a first order these two expressions are the same, so full optimization of hours worked only leads to second order utility gains approximately of size

$$U(c_\varepsilon^*, h_\varepsilon^*) - U(c_\varepsilon^*, h^*) \simeq \frac{1}{2}\gamma(1+\gamma)\varepsilon^2. \tag{12.35}$$

Now, we introduce optimization frictions through the fixed utility loss Δ. In order for the individual to choose to change hours in order to respond to the wage change, Δ must be smaller than (12.35), or

$$\varepsilon > \varepsilon_{\min} = \left(\frac{2\Delta}{\gamma(1+\gamma)}\right)^{\frac{1}{2}}.$$

Note that the higher the labor supply elasticity, the lower this threshold is because, when γ is very large, even small changes in hours lead to large changes in utility, as apparent from (12.34). Thus, in the presence of fixed cost of re-optimization, hours might not be responsive to small changes in wages, which creates a downward bias in the estimation of the structural parameter γ. However, hours would respond with elasticity γ to large enough changes in wages. Specifically, in the presence of optimization frictions of size Δ, it descends

from (12.33) that we would only observe changes in log hours worked (approximately) larger than

$$\Delta \log h^* > \gamma \varepsilon_{\min} = \left(\frac{2\Delta\gamma}{1+\gamma} \right)^{\frac{1}{2}}.$$

Taking stock, there might be several frictions that induce agents to deviate from the optimal choices predicted by standard frictionless economic models. When these frictions are salient, microeconometric studies of labor supply can be uninformative, unless they analyze episodes where wage changes are large enough to overcome these frictions. We conclude by noting that this is a more general insight that applies beyond labor supply.

12.5 Labor supply and balanced growth

As illustrated in Chapter 13, developed economies tend to grow at a constant rate, and this growth path is "balanced,"meaning that consumption, investment and output grow at the same rate. Hours worked, however, cannot keep growing at a constant rate because, inevitably, at some point they would hit the time endowment and stop growing, violating balanced growth. Figure 12.1 shows that in the U.S. hours per person have remained relatively stable over time. King, Plosser, and Rebelo (1988) showed that this property of long-run labor supply puts important restrictions on preferences.

Consider the social planner formulation of the neoclassical growth model where output is produced by a Cobb-Douglas production function with constant capital share α (another property of balanced growth):

$$\max_{\{C_t, H_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U\left(C_t, 1 - H_t\right) \tag{12.36}$$

subject to

$$C_t + K_{t+1} = Z_t^{1-\alpha} K_t^{\alpha} H_t^{1-\alpha} + (1-\delta) K_t$$

and

$$K_0 \text{ given.}$$

Let Z_t denote labor-augmenting technological change which we assume to grow at a constant rate g, i.e. $Z_{t+1} = Z_t (1+g)$ for all t. Along a balanced growth path, consumption, investment, capital and output all grow at rate g and hours are constant. What do these features imply for U? In Appendix 12.A we prove that the period utility function U is consistent with balanced growth—in particular with the observation that hours worked are constant as the economy grows at a constant rate—if and only if

$$U(C, 1-H) = \begin{cases} \dfrac{C^{1-\sigma} - 1}{1-\sigma} v\left(1-H\right) & \text{if } \sigma \neq 1 \\ \log\left(C\right) + v\left(1-H\right) & \text{if } \sigma = 1. \end{cases} \tag{12.37}$$

In a static setting with only labor income, these specifications have the implication that a higher wage rate has no effect on desired hours of work, i.e., income and substitution effects exactly cancel out. This is what allows hours worked to remain constant along the balanced

growth path in an economy with positive productivity growth, which in turn implies growth in real wages over time.

Figure 12.1 also shows that, for many countries, hours worked have slowly diminished over the post-war period. Thus, the postwar U.S. experience, over which hours have shown no net decrease and which is the main argument for the use of "balanced-growth preferences"(12.37), is to some extent a striking exception more than a representative feature of modern economies. In the next section we discuss factors that can account for the variation in hours worked across countries and over time.

12.6 Labor supply across countries and over time

Two classic issues in the study of labor supply are that (i) there exist large differences in hours worked per capita across countries, as shown in Table 12.1, and (ii) hours worked vary widely over time at the business cycle frequency, as shown in Figure 12.4. This section serves to illustrate how representative household models have been used to shed light on the role of labor supply in each of these contexts.

A key message from this section is that the optimal labor supply implied by a representative agent model with a simple and parsimonious specification of preferences–a period utility function of the form

$$U(c_t, h_t) = \log(c_t) - \frac{\psi}{1+1/\gamma} h_t^{1+1/\gamma}$$

accounts for an important part of variation in aggregate hours of work, both across countries and over time. Put somewhat differently, labor supply is relevant for understanding aggregate labor market outcomes. While quantitatively important, this parsimonious model of labor supply does not capture all aspects of the data, indicating that some of the variation found in the data reflects factors other than labor supply.

12.6.1 Labor supply and taxation across countries

A variety of factors can explain cross-country differentials in hours worked. A chief candidate is the level of taxes and transfers because they strongly influence work incentives. Prescott (2004) shows that households work less in countries with high tax rates on labor, and develops a simple theoretical framework to assess this hypothesis quantitatively.

Consider the problem of an infinitely-lived representative household who faces tax rates τ_c and τ_h on consumption and labor income, and receives $\mathcal{T}_t$ as lump-sum transfers:

$$\max_{\{C_t, H_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log C_t - \psi \frac{H_t^{1+1/\gamma}}{1+1/\gamma} \tag{12.38}$$

subject to

$$(1+\tau_c) C_t + K_{t+1} = (1-\tau_h) w_t H_t + r_t K_t + (1-\delta) K_t + \mathcal{T}_t.$$

A representative firm produces the final good with a Cobb-Douglas technology so the resource constraint for this economy is

$$C_t + G_t + K_{t+1} - (1-\delta) K_t = Y_t = Z_t^{1-\alpha} K_t^{\alpha} H_t^{1-\alpha},$$

where G_t is government consumption. The government uses tax revenue to finance spending on government consumption G_t and the lump sum transfer $\mathcal{T}_t$, subject to a period by period balanced budget. Note that the specification of preferences assumes that households do not value government consumption.[11]

Prescott assumes Z_t grows at a constant rate and $G_t = gY_t$ and focuses on the level of hours worked along the balanced growth path. Starting from the first order condition for the optimal intratemporal allocation of consumption one can derive the following expression for hours along the balanced growth path:

$$H^* = \left[\frac{(1-\tau)(1-\alpha)}{\chi\psi}\right]^{\frac{\gamma}{1+\gamma}} \tag{12.39}$$

where

$$\tau = \frac{\tau_h + \tau_c}{1+\tau_c}$$

is the effective tax return to labor income and $\chi \equiv C_t/Y_t$ is the (constant) consumption-output ratio along the balanced growth path.

Prescott takes the view that the capital share α, the labor disutility shifter ψ, and the Frisch elasticity γ are common across countries. Under this assumption, the only two terms in (12.39) that can explain cross-country differences in hours worked are τ and χ.[12]

As a concrete example, consider the U.S. and Italy. Prescott reports that in the mid 1990s, the average American aged 15-64 worked 25.9 hours per week, while the average Italian worked only 16.5 hours per week. The ratio between the two is a staggering 1.57. Prescott reports that, for the mid 1990s, τ equals 0.40 in the U.S. and 0.64 in Italy, and χ is 0.81 in the U.S. and 0.69 in Italy. If we set $\gamma = 1$, i.e. a unitary aggregate Frisch elasticity, we obtain a ratio of hours worked between the U.S. and Italy predicted by the theory of 1.42 compared to 1.57 in the data. Calculations for a number of other countries confirm that this simple theory can account for much of the cross-country differences in labor supply.

Recall our earlier discussion in which we argued that the elasticity that matters for permanent changes in taxes is either the Marshallian when the tax revenue is not rebated, or the Hicksian when it is rebated. Prescott set $\sigma = 1$ based on balanced growth path considerations. Conditional on fixing the value of σ, both of these elasticities are determined by the value of γ, which is why the key elasticity parameter in equation (12.38) is γ.

The previous discussion focused on differences in hours worked across a sample of advanced economies. Bick, Fuchs-Schündeln, and Lagakos (2018) studied differences in hours across a broader sample of countries and found a systematic negative relationship between hours of work and the level of development as measured by GDP per capita. While it is true that less developed economies tend to have smaller tax and transfer systems than advanced economies, Bick et al. (2018) argue for an alternative explanation: the reason that hours of

[11]The results of the analysis are unaffected if one alternatively assumes that individuals value G_t but it enters utility separably with respect to C_t and H_t.

[12]Variation in χ in the data can come from variation in g or variation in the ratio of investment to output. The model as specified does not generate variation in the investment to output ratio. As a practical matter, variation in investment to output is relatively unimportant in this context. We note that variation in capital income tax rates is one possible source of variation in the investment to output ratio.

work are higher in poorer economies is because income effects are larger than substitution effects.

As discussed earlier, a commonly used specification of the period utility function that displays perfectly offsetting income and substitution effects is the following:

$$U(c_t, h_t) = \log(c_t) - \frac{\psi}{1 + 1/\gamma} h^{1+1/\gamma}.$$

There are two modifications to this specification that are commonly used to capture the possibility that income effects are larger than substitution effects. The first introduces a subsistence consumption term $\bar{c}$, resulting in a period utility function given by:

$$U(c_t, h_t) = \log(c_t - \bar{c}) - \frac{\psi}{1 + 1/\gamma} h^{1+1/\gamma}.$$

With this specification, individuals need to work enough hours in order to reach consumption of at least $\bar{c}$.[13] It follows that very low wages will lead individuals to choose to work long hours in order to reach consumption of at least $\bar{c}$. A key feature of this specification is that the magnitude of the income effect relative to the substitution effect decreases with the level of consumption, with the two effects being perfectly offsetting in the limit as consumption becomes very large.

A second specification that allows for the income effect to be larger than the substitution effect is:

$$U(c_t, h_t) = \frac{1}{1 - \sigma} c_t^{1-\sigma} - \frac{\psi}{1 + 1/\gamma} h_t^{1+1/\gamma},$$

where $\sigma > 1$. Unlike the previous specification with a subsistence consumption term, a key feature of this specification is that the income effect is always larger than the substitution effect. Boppart and Krusell (2020) show that this specification gives rise to a balanced growth path in which hours of work decrease at a constant rate. They argue that such a balanced growth path offers a better description of long run changes in hours worked among current advanced economies. Ohanian, Raffo, and Rogerson (2008) offer an alternative interpretation of this time series evidence. They argue that changes in tax rates over time can account for much of trend behavior of hours among OECD economies after 1960. Whether income effects dominate substitution effects, and if so, whether this effect diminishes as consumption grows, remains an open issue.

12.6.2 Labor supply and business cycle fluctuations

When discussing the Frisch elasticity, we have stated that this elasticity is especially relevant to understand the role of labor supply in accounting for the large fluctuations of hours worked over the business cycle. Figure 12.4 shows that hours worked are very volatile over the cycle and are pro-cyclical. To illustrate the role of the Frisch elasticity for aggregate fluctuations

[13] Our discussion here implicitly assumes that it is feasible for the individual to achieve consumption greater than $\bar{c}$. More generally, one would need to modify this specification to define utility when consumption is less than $\bar{c}$.

in hours, it is useful to return to the household problem in (12.38), allowing now productivity Z_t to be stochastic and to be the driver of economic fluctuations. Equation (12.39) which determines optimal hours worked still holds, but now the consumption-output ratio varies over time. Expressed in logarithms, equation (12.39) yields

$$\log H_t = const - \frac{\gamma}{1+\gamma} \log \left(\frac{C_t}{Y_t} \right), \tag{12.40}$$

where *const* is a constant term. Expectations of changes in productivity are all captured by the current consumption-output ratio. Suppose productivity is random and is currently above trend, i.e. the economy is in a boom, but expected to revert towards the trend, for example because it follows an AR1 process. The permanent income hypothesis suggests that consumption should increase less than income, as some of this increase in income is saved (see Chapter 11). Then, the consumption-output ratio will be low, and labor supply will be above its steady-state value. The sensitivity of aggregate hours worked H_t to fluctuations in aggregate productivity, and hence in the consumption-output ratio, is mediated by the value of the aggregate Frisch elasticity γ.

If the Frisch elasticity is infinite (as implied by the indivisible labor model of Rogerson (1988) and Hansen (1985) discussed later in this chapter) then the ratio $\gamma/(1+\gamma)$ equals one and log hours move one for one with changes in $\log(C_t/Y_t)$. But notably, even a Frisch elasticity of 1, a value consistent with the empirical evidence as we will argue later in Section 12.8, produces a response which is already half the size of the infinite-elasticity case. There is broad agreement that the Frisch elasticity is central to the business cycle behavior of aggregate labor supply.

We note that, in absence of capital and in a closed economy, $C_t = Y_t$ at every t, thus these preferences consistent with balanced growth imply that hours worked are invariant to any productivity change, including temporary ones. Since the procyclicality of aggregate hours is a salient stylized fact of developed economies, it is important for macroeconomic models that aim to be consistent with this key feature of the data to include capital and savings decisions.

Before concluding we should point out that, when taken to the data, equation (12.40) yields, typically, a poor approximation of aggregate hours dynamics at the business cycle frequency. In particular, hours decrease more than what would be predicted by movements in the consumption-output ratio during recessions. The time-varying term missing from equation (12.40) is called the *labor wedge* and acts like a countercyclical labor income tax in the intratemporal first-order condition of the representative agent. Chapter 14 discusses a number of possible interpretations of the labor wedge.[14] The most plausible explanation is that the neoclassical labor supply model which underlies (12.40) abstracts from labor market frictions, bilateral monopoly power in wage setting, and unemployment, all factors that can affect that relationship. Chapter 20 is devoted to these topics.

[14] Shimer (2009) provides a critical overview of such interpretations.

12.7 Dynamic returns to labor supply

The analyses of labor supply described so far all assume wages evolve exogenously over the life cycle. There is a long tradition in macroeconomics that takes a more plausible view that wages increase when individuals accumulate work experience or human capital. Imai and Keane (2004) show that explicitly incorporating human capital accumulation into a life-cycle model has sharp implications for the measurement of labor supply elasticities.

To focus on human capital, we abstract from borrowing constraints in the life-cycle problem (12.7), but we add dynamic returns to work. The individual problem becomes:

$$\max_{\{c_{it},h_{it}\}_{t=0}^{T}} \mathbb{E}_t \sum_{t=0}^{T} \beta_i^t \left[\frac{c_{it}^{1-\sigma}}{1-\sigma} - \psi \frac{h_{it}^{1+1/\gamma}}{1+1/\gamma} \right]$$

subject to

$$c_{it} + a_{i,t+1} = Ra_{it} + (1-\tau)\, w_{it} h_{it} + \mathcal{T}$$

and

$$w_{it} = \left(1 + \kappa \sum_{j=0}^{t-1} h_{ij} \right) w_{i0}.$$

The second constraint is the new feature of the model. It states that the hourly wage at time t depends on all past hours worked from $j = 0$ to $t-1$. Thus, in this model, the return to an hour of work is not just the current wage as in (12.7). It consists of the current wage plus the expected present value of increased earnings in all future periods because working an additional hour today raises the wage permanently from today onward. Imai and Keane refer to this second component of the total return to work as the human capital term.

The first order condition with respect to hours worked at date t is given by

$$\psi h_{it}^{\frac{1}{\gamma}} = \lambda_{it} (1-\tau) \left[w_{it} + \underbrace{\kappa w_{i0} \sum_{j=1}^{T-t} R^{-j} \mathbb{E}_t \left[h_{i,t+j} \right]}_{\text{human capital}} \right] \tag{12.41}$$

where we have used the Euler equation to simplify the right-hand side of (12.41). This optimality condition illustrates that the human capital term (the second term in the right-hand side) creates a wedge between the current wage and the return to work. As a consequence, γ no longer captures the Frisch elasticity. The parameter γ is the elasticity of hours worked to the total return to work, whereas the Frisch elasticity measures the impact of an anticipated or transitory change in w_{it} on h_{it}, i.e., keeping the marginal utility of wealth constant.

To understand why the presence of dynamic returns to work can affect the estimation of γ and of the intertemporal labor supply elasticity, consider Figure 12.6 adapted from Keane and Rogerson (2015). This figure plots the life cycle paths of hours, wages, the human capital term, and the return to work. Note that the human capital term is high early on, but it declines with age. The reason is that the return to working more in order to increase future wages is larger when the individual is young, i.e. t is small, as clear from the sum in equation

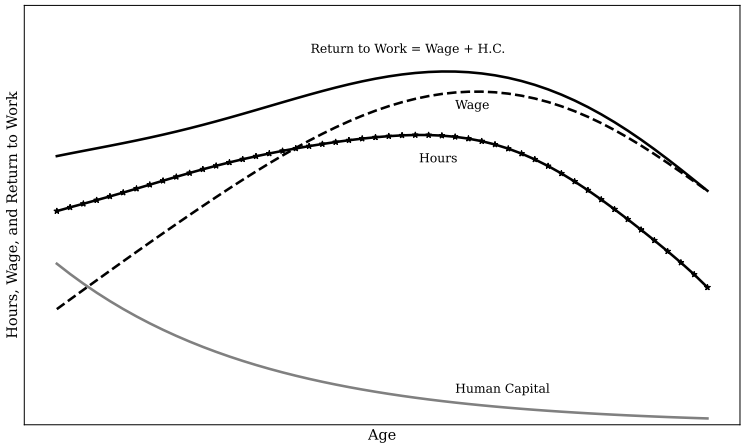


Figure 12.6: Life cycle profile of hours, wages, the human capital term, and the return to work in the Imai-Keane model

(12.41) because the worker can enjoy the higher wages for longer. Ignoring these dynamic returns and naively estimating γ from the relation between current wages and current hours would lead one to underestimate its true value because, over the first half of the life-cycle, hours are relatively flat while wages grow steeply. Instead, γ should be gauged from the joint dynamics of hours and the return to work, whose path is much flatter because the profiles of hourly wages and human capital tend to offset each other.

Turning to the Frisch elasticity, it is clear that the larger is the gap between the return to work and the current wage, i.e., the human capital term, the smaller the Frisch elasticity will be, since the optimal decision on how much to work at time t depends both on current and dynamic returns. Increasing the current return w_{it} may not have a large impact on h_{it} if the dynamic returns are large. In line with this logic, Imai and Keane estimate a Frisch elasticity that increases with age from 0.4 to 2.0.

Another interesting implication of the human capital model is that it can alter the ranking of labor supply elasticities. Recall that in Section 12.4 we showed that in the benchmark dynamic model, Frisch > Hicksian > Marshallian. This ordering is no longer necessarily true here. To understand this point, compare a transitory tax increase that only affects today's after-tax earnings and a permanent tax increase that affects all future disposable earnings. Obviously, the income effect which pushes in the direction of working longer hours is more powerful for a permanent tax hike. However, the substitution effect will also be stronger. The reason is that a permanent tax increase reduces the total return to work —both the current wage and the human capital term. Under some parameterizations, the substitution effect can dominate the income effect. Thus, although human capital dampens labor supply responses to transitory tax changes (and this effect is especially strong for young workers), it magnifies the impact of permanent changes.

12.8 Labor supply along the intensive and extensive margins

In Section 12.2, we highlighted that variation in aggregate hours of work are accounted for by variation along both the extensive (number of people employed) and intensive (hours per person employed) margins. The models presented in the previous sections of this chapter have the feature that all of the variation in aggregate hours occurs along the intensive margin, with no consideration of the extensive margin. A large body of work has shown that explicitly modeling both margins of adjustment is important for understanding labor supply responses. In this section we study models that incorporate intensive and extensive margins and illustrate some of their implications. One of the key messages is that labor supply elasticities estimated using individual panel data may not be informative about aggregate labor supply elasticities, i.e., how aggregate hours of work respond to changes in the average real wage.

12.8.1 A static model with indivisible labor

We begin by studying a static model that forces all adjustment to be along the extensive margin and contrasting its properties with those of a model in which all adjustment takes place at the intensive margin. The intensive margin economy will essentially be a general equilibrium version of the benchmark static model introduced in Section 12.3. In particular, we assume a unit mass of identical individuals, each with preferences of the form:

$$U(c, h) = u(c) - \psi v(h)$$

with u and v taking on the same functional forms as in Section 12.2:

$$u(c) = \frac{1}{1-\sigma} c^{1-\sigma}, \quad v(h) = \frac{1}{1+1/\gamma} h^{1+1/\gamma}.$$

There is a constant returns to scale aggregate production function that uses only labor:

$$Y = AH$$

where H is total labor input. Individuals in this economy are allowed to choose any value for their hours, but for reasons that will become clear we will normalize the total discretionary time of each individual to equal one, so that h must lie between 0 and 1.

The extensive margin only economy is identical with one exception: instead of allowing each individual to freely choose their hours of work, we now limit individuals to two choices: either $h = 0$ or $h = \bar{h}$, where $\bar{h}$ is a positive constant lying strictly between 0 and 1. This restriction can be interpreted as capturing the fact that many jobs come with a standard workweek, so that the key choice that an individual makes is whether to work at the standard workweek or to not work. Because labor can only be supplied in the fixed quantity $\bar{h}$, this model is referred to as the indivisible labor model; the economy in which individuals can freely choose h will be referred to as the divisible labor model.

As a first step, we examine what a social planner would do in each of these economies. In the divisible labor model, it is easy to show that the Social Planner will give all individuals the same allocation of consumption and hours. These values are the solution to:

$$\max\ U(c,h)$$

subject to

$$c = Ah,\ c \geq 0, \text{ and } 0 \leq h \leq 1.$$

In the indivisible labor economy, the social planner's problem is slightly more complicated. By assumption, the social planner can only give all individuals the same allocation of consumption and hours if either all individuals work $\bar{h}$ or all individuals do not work at all. But in general, the social planner may want to have only a fraction of all individuals work. Denote this fraction by e. Conditional on having a fraction e of individuals employed, the social planner has to decide how to allocate the resulting output of $Ae\bar{h}$ among all of the individuals in the economy. Because preferences are separable and the function u is strictly concave, it is optimal for the social planner to spread consumption equally across all individuals, independently of whether they are chosen to work. An alternative to having the social planner treat identical individuals asymmetrically is for the social planner to offer all individuals the same random allocation, namely, that they work with probability e and receive the same consumption c independently of the realization of the randomness. Formulated this way, the social planner in the indivisible labor economy will solve:

$$\max_{e,c}\ eU(c,\bar{h}) + (1-e)U(c,0)$$

subject to

$$c = A\bar{h}e,\ 0 \leq e \leq 1, \text{ and } c \geq 0.$$

Note that we are implicitly appealing to the law of large numbers; given that there is a unit mass of individuals, if each has a probability e of working then we know that a fraction e of workers will be selected to work.

Next we examine the objective function of the social planner in the indivisible labor economy in more detail. Recalling that U is assumed to be separable we have:

$$\begin{aligned} eU(c,\bar{h}) + (1-e)U(c,0) &= u(c) - ev(\bar{h}) - (1-e)v(0) \\ &= u(c) - e[(v(\bar{h}) - v(0)] - v(0). \end{aligned}$$

The term $v(0)$ is just a constant, and adding a constant to an objective function does not affect the optimal choices, so that the objective function is equivalent to:

$$u(c) - \tilde{\psi}e$$

where $\tilde{\psi} = v(\bar{h}) - v(0)$. Compare this to the objective function of the social planner in the divisible labor economy, which is:

$$u(c) - \frac{\psi}{1+1/\gamma}h^{1+1/\gamma}$$

Recalling our discussion of the Frisch elasticity in the simple two period example earlier in this chapter, we see that whereas the divisible labor economy has an aggregate Frisch elasticity equal to γ, the indivisible labor economy has an infinite aggregate Frisch elasticity. The indivisible labor economy highlights the need to distinguish between individual and aggregate elasticities. In the divisible labor economy, all individuals work the same amount and movements in aggregate labor supply are exactly mirrored in movements in individual labor supply. But in the indivisible labor economy, a given individual either works or does not work, so individual labor supply does not vary smoothly. Nonetheless, aggregate labor supply does vary smoothly with changes in the fraction of individuals that work.

In summary, this analysis of indivisible labor offers two key takeaways. First, it creates a disconnect between the properties of aggregate labor supply and individual labor supply. Second, it explains why aggregate labor supply might have a very large Frisch elasticity independently of the Frisch elasticity that each individual has.

12.8.2 Decentralizing the social planner's allocations

Our discussion so far has focused on the allocations that would be chosen by a social planner. In the divisible labor economy it is straightforward to show that the social planner allocation would be implemented if we studied a competitive equilibrium. In the indivisible labor economy we argued that a social planner would use randomization. It is not immediately obvious how to implement such allocations as competitive equilibria. Following Hansen (1985) and Rogerson (1988), we briefly describe how this can be done if we allow individuals to sell employment probabilities. If w is the wage per unit of work, then offering to work $\bar{h}$ hours with probability e will result in labor earnings of $e\bar{h}w$. The firm demands a certain number of total work hours, which must be equal to the total supply of labor in equilibrium. But appealing to the law of large numbers, this is not an issue since aggregate labor supply will be deterministic even if each individual offers a random amount of hours.

While the above discussion shows how one can formally define a market structure such that the social planner's allocation is implemented as a competitive equilibrium, one might argue that this market structure with employment lotteries does not seem to be a good description of the market structure we observe in real-world economies. Ljungqvist and Sargent (2006) have shown that the social planner's allocation can be implemented as a competitive equilibrium with a more compelling market structure. Their argument requires explicit consideration of an economy with many time periods. To make the argument precise we will focus on an economy in continuous time that runs from time 0 to time T. At each instant of time this economy resembles the static indivisible labor economy that we described above, and we assume that all parameters are constant over time. For simplicity, we assume that individuals evaluate lifetime utility as the integral of utility over time without any discounting.

If the social planner's optimal allocation in the static economy was for everyone to work with probability e^* and receive consumption c^*, then it follows that an optimal allocation in the dynamic economy will be to choose these values at each point in time. The key insight of Ljungqvist and Sargent is that from a lifetime perspective, choosing to work with probability e at each instant of time is equivalent to deterministically choosing to work a fraction e of time. If an individual chooses to deterministically work in a fraction e of the time periods,

there will be some instants in which they have labor income equal to $w\bar{h}$ and other instants in which they have labor income equal to zero. But if they have access to markets that allow for borrowing and saving, and the interest rate is equal to zero, they can achieve a smooth profile for consumption equal to $e\bar{h}w$ in each period. Because we assumed that individuals do not discount future utility, it turns out that the competitive equilibrium interest rate will be equal to zero. Additionally, if we normalize the price of consumption to equal unity at each instant, the equilibrium wage rate will be constant and equal to A. In the competitive equilibrium each individual solves the following problem (a continuous-time formulation is adopted so that we can avoid the integer constraint):

$$\max \int_0^T U(c(t), e(t)\bar{h})dt$$

subject to

$$\int_0^T c(t)dt = \int_0^T e(t)\bar{h}Adt,\ e(t) \in \{0, 1\},\ \text{and}\ c(t) \geq 0.$$

The solution to this problem is such that the individual wants to choose the $e(t)$ such that $\int_0^T e(t)dt = e^*$ and $c(t) = c^*$ but is indifferent about at which instant they work. That is, they care about how much they work over their lifetime but are indifferent about the timing of that work. Because this is true for all individuals it is possible to have a fraction e^* work in each period, so that the social planner's allocation can be implemented in competitive equilibrium without having any markets for random supply of labor.

12.8.3 Allowing for heterogeneity

Our previous analysis argued that an indivisible labor economy with identical individuals gives rise to an aggregate Frisch elasticity that is infinite if the aggregate employment rate is interior.[15] In this subsection we argue that this result hinges on the assumption of identical individuals. This is intuitive. In an economy in which everyone is identical, their reservation wage, i.e., the wage at which they are indifferent between working and not working, is the same for all individuals. In such a situation a small change in the wage around the reservation wage can move everyone from wanting to work to not wanting to work and vice versa. In contrast, if there is heterogeneity in reservation wages then a small change in the wage rate will only change the decisions of a small group of people. We now develop this idea more formally in one specific context.

For simplicity we return to a static economy. We assume that individuals are identical ex ante, but are subject to idiosyncratic shocks to their disutility of working parameter ψ. In particular, each individual will receive an iid draw from a distribution with cdf $F(\psi)$ and density $f(\psi)$. The aggregate production function is the same as before.

We again focus on the social planner's problem for this economy. Because all individuals are the same ex ante, we assume that the social planner maximizes the equal weighted integral of expected utility across individuals. The social planner will choose individual

[15] If the social planner chooses $e^* = 1$ then by continuity a small change in the wage at any instant holding all other wages constant will not result in any change in e^*, implying an aggregate Frisch elasticity equal to zero.

allocations contingent on the realizations of disutility of working, and can be written as two functions, $e(\psi)$ and $c(\psi)$, which represent the work probability and consumption allocation respectively for an individual with realization ψ. With separable utility, the social planner will continue to allocate the same consumption to all individuals independently of their disutility of working. For the employment decision, the most efficient way to have a fraction e of all individuals work is to choose the fraction e of individuals with the lowest draw for the disutility of work and to have them work with probability one, while the other individuals work with probability zero. This amounts to choosing a reservation disutility level ψ^* with the property that $F(\psi^*) = e$. Every choice of e is associated with a unique choice of ψ^*, which we will write as $\psi^*(e)$. It then follows that the total disutility associated with an employment rate of e, which we denote by $\tilde{v}(e)$ is given by:

$$\tilde{v}(e) = \int_0^{\psi^*(e)} \psi f(\psi) d\psi$$

and the social planner's problem can be written as:

$$\max \; u(c) - \tilde{v}(e)$$

subject to

$$c = Ae\bar{h},\ 0 \leq e \leq 1, \text{ and } c \geq 0.$$

The key point is that the function $\tilde{v}$ is no longer linear; as e increases the social planner will be choosing workers with progressively higher values of ψ to work. The aggregate Frisch elasticity will now depend on the local properties of the $\tilde{v}$ function, which will in turn depend on the local properties of the density function f. This clearly illustrates that the aggregate Frisch elasticity depends critically on the distribution of heterogeneity in the population.

The previous discussion considered a static model, but one can easily generalize it to a dynamic economy by assuming that individuals receive iid draws of ψ at each instant from the distribution $F(\psi)$. As in the previous subsection, the social planner's solution can be implemented in a competitive equilibrium without any insurance markets if we allow complete markets for borrowing and lending. Our discussion has focused on heterogeneity in preferences, but similar effects are obtained if one considers heterogeneity in productivity.

A key question is to assess the implications of indivisible labor when allowing for empirically relevant sources of heterogeneity. In particular, what does such a model imply for the values of individual and aggregate labor supply elasticities. One of the first quantitative analyses of this question was undertaken by Chang and Kim (2006).[16] They consider an aggregate model populated by infinitely-lived two-member households in which labor supply is indivisible, individuals are subject to idiosyncratic productivity shocks and markets for credit and insurance are incomplete. Consistent with the unitary household model that we described in Section 12.3.2, households consist of a male and a female, and the household has preferences given by:

$$\sum_{t=0}^{\infty} \beta^t \left[2 \log\left(\frac{c_t}{2}\right) - \psi_m \frac{h_{mt}^{1+1/\gamma}}{1+1/\gamma} - \psi_f \frac{h_{ft}^{1+1/\gamma}}{1+1/\gamma} \right]$$

[16] Additional issues were explored in Chang and Kim (2007) and An, Chang, and Kim (2009).

where c_t is household consumption, and h_{mt} and h_{ft} are hours worked by the male and female household member. Each individual can only supply 0 or $\bar{h}$ units of labor in any period. Individual productivity, denoted by z_t, is stochastic and follows the stochastic process:

$$\log z_{jt+1} = \rho_j \log z_{jt} + \varepsilon_{jt+1},\ j = m, f. \tag{12.42}$$

The process is gender specific but is the same for all individuals of a given gender. The realizations of the ε_{jt} are iid across individuals and over time. A worker with productivity z_t has labor earnings $w_t z_t \bar{h}$ if working, where w_t is the wage per efficiency unit of labor in period t.

The production side of the economy is the same as in the standard growth model. There is a Cobb Douglas aggregate production function that uses capital and efficiency units of labor, output can be used as either investment or consumption, and capital depreciates at a constant rate δ.

Chang and Kim study a competitive equilibrium assuming the following market structure. Each period there are markets for capital and labor services as well as output, but the market for labor services does not allow workers to sell employment probabilities. Following the work of Aiyagari (1994) and Huggett (1993a) discussed in Section 11.4, markets for credit and insurance are incomplete. Specifically, there are no markets for insurance against idiosyncratic productivity shocks, and households are able to save and borrow in a credit market, subject to an exogenous borrowing limit.

Chang and Kim calibrate this model assuming that a period is equal to one quarter and show it matches the cross-sectional heterogeneity in earnings and wealth found in the data reasonably well, with the exception that it is not able to capture the extreme right tail of the wealth distribution. The model implies that individuals move between employment and nonemployment as their productivity and asset holdings evolve. Chang and Kim did not assess the extent to which the transitions in the model match those found in the data, but subsequent related work has shown this to be the case.[17]

They proceed to study the properties of individual and aggregate labor supply in their calibrated model. First, they simulate histories lasting 120 quarters for a sample of households in the steady state, aggregate the observations to produce a panel data set at annual frequency and then run the following panel regression using individuals with positive hours in each year:

$$\log h_{it} = \gamma(\log w_{it} - \log c_{it}) + \varepsilon_{it}. \tag{12.43}$$

Recall that in Section 12.3 we noted that this equation can be derived in the context of a two period setting if the function u is given by log. Their estimation exercise yields estimates for γ of .41 and .78 for males and females respectively. The key finding is that a standard labor supply regression using individual data generated by their model yields a relatively small estimate of the Frisch elasticity for men, and a larger and moderate estimate for women.

Their second exercise follows in the spirit of the literature on real business cycles. In particular, they now assume an aggregate technology shock that follows an AR(1) process

[17] Bils, Chang, and Kim (2012) examine this in a slightly more general model. Krusell, Mukoyama, Rogerson, and Şahin (2017) further extend this analysis by including search frictions and considering movements between employment, unemployment and out of the labor force, though they consider single individual households.

and simulate the economy to produce aggregate time series data for hours, consumption and wages, and then run the same regression as above but now using aggregate time series data. This produces an estimate for γ of 1.08. Notably, this estimate is higher than both of the estimates obtained when using individual data.

Lastly, they consider a representative household model with preferences of the form:

$$\sum_{t=0}^{\infty} \beta^t \left[\log(c_t) - \tilde{\psi} \frac{h_t^{1+1/\tilde{\gamma}}}{1+1/\tilde{\gamma}} \right],$$

where h_t is now allowed to take on any value in the interval $[0, 1]$. Assuming the same process for aggregate technology shocks, they solve for the value of $\tilde{\gamma}$ that generates fluctuations in aggregate hours that are the same as in the heterogeneous agent economy with indivisible labor, and find a value for $\tilde{\gamma}$ of approximately 2. This value is roughly five times as large as the Frisch elasticity estimated from individual data on males.

Their analysis has two key messages. First, allowing for empirically reasonable individual heterogeneity and risk-sharing in a model with indivisible labor dramatically lowers the implied aggregate Frisch elasticity, reducing it from infinity to around 2. But second, the model still implies a large disconnect between the labor supply elasticity obtained using standard methods on microdata for continuously employed individuals and the aggregate elasticity.

12.8.4 Models with intensive and extensive adjustment

So far in this section we have focused on models in which all adjustment in aggregate hours occurs either along the intensive margin or along the extensive margin. As emphasized in the first section of this chapter, data shows that there is important adjustment along both margins. In this subsection we develop some models that feature adjustment along both margins. The key feature that we use to generate this outcome is to assume a convex mapping from the hours supplied by an individual to the efficiency units of labor associated with these hours. There are two specifications of this form that are commonly found in the literature. One is to assume that there is a fixed time cost associated with working:

$$h^e = \max(0, h - h_f), \tag{12.44}$$

where h is the time devoted to work, h^e denotes the efficiency units of labor and $h_f > 0$ is the fixed time cost associated with working. The key distinction is that h is what enters the utility function of the individual while h^e is what enters the production function. The fixed cost h_f can represent some combination of the time associated with getting to and from work and the time required to get set up once at work. The presence of fixed costs makes it suboptimal for workers to choose to work low numbers of hours and creates an incentive to concentrate hours in order to minimize the total fixed costs.

The second specification assumes a smooth convex mapping from hours of work to efficiency units of labor supplied:

$$h^e = h^\theta,$$

where $\theta > 1$ ensures that this relationship is convex, and $\theta = 1$ represents the standard textbook case in which time devoted to work is the same as the efficiency units of work.

This specification is consistent with a wage penalty for part-time work and a wage bonus for overtime work. Similar to the fixed-cost specification, when $\theta > 1$ there is an incentive for an individual to concentrate their hours of work, since for example, an individual who works 25 hours in each of two periods will earn higher income if they instead chose to work 50 hours in one of the periods and 0 hours in the other period.

Because both of these specifications incentivize individuals to concentrate hours of work and avoid periods in which they work relatively few hours, they both have a force that captures the essence of the indivisible labor assumption. In fact, if we adopt either of these specifications in an economy with identical workers, then for sufficiently large values of h_f or θ the solution to the social planner's problem will look like the outcome in an indivisible labor economy, with one group of individuals working $\bar{h}$ units and the remaining individuals all working 0 hours. But importantly, the value of $\bar{h}$ will be endogenously determined, depending upon the primitives of the model, and in particular the value of h_f or θ.

When individuals are heterogeneous, these specifications lead to labor allocations in which variation in aggregate hours reflects changes both in the fraction of individuals employed and hours per employed person. Rogerson and Wallenius (2009) illustrated this in the context of a life cycle model to examine the effects of tax and transfer programs. They used the fixed time cost of work specification and introduced heterogeneity by assuming that productivity varied in a deterministic manner over the life cycle. In particular, assume that time is continuous, an individual lives from $t = 0$ to $t = T$, and that productivity varies deterministically over the life cycle according to the function $z(t)$. The individual maximizes lifetime utility, which is given by (again, a continuous-time formulation is employed to avoid the integer constraint):

$$\int_0^T [u(c(t)) - v(h(t))]dt,$$

where for simplicity we assume no discounting. As above, the individual faces the mapping from hours of work h to efficiency units h^e as specified in equation (12.44). The wage per efficiency unit of labor supply is constant over time and equal to w, and the individual is allowed to borrow and save at the interest rate of 0 subject to the constraint that they they do not die with any debt. That is, they face a lifetime budget constraint given by:

$$\int_0^T c(t)dt = w \int_0^T \max(0, h(t) - h_f)dt.$$

To make things concrete, assume that z is continuous, has a single peak, and that $z(0) = z(T) = 0$. The optimal profile for life cycle hours in this model has the following form. There is a threshold value z^* for individual productivity such that when productivity is below z^* the individual chooses $h = 0$. When productivity is greater than z the individual chooses to work positive hours, and hours are an increasing function of productivity. Thus, at the individual level, lifetime labor supply features adjustment along both the intensive and extensive margins.

Rogerson and Wallenius use this framework to quantitatively assess its implications for some labor supply elasticities. For their quantitative work they adopt $u(c) = \log c$ and $v(h) = \psi h^{1+1/\gamma}/(1 + 1/\gamma)$, and assume that life-cycle productivity $z(t)$ is piecewise linear. They consider values of γ ranging from .10 to 2.00 and in each case calibrate the model to match some key properties of life cycle labor supply.

Two key results emerge. First, they assess the response of aggregate hours to a permanent tax and transfer policy that includes a balanced budget constraint and find that the response is roughly independent of the value of γ. Specifically, increasing the tax rate from 30% to 50% decreases aggregate hours by roughly 20% for all values of γ in the range of 0.10 to 2.00. Recall from our previous discussion that this policy exercise reflects the value of Hicksian elasticity of labor supply.

Second, although γ has virtually no effect on the change in aggregate hours, it does have a quantitatively important effect on the relative importance of adjustment along the intensive and extensive margins. For example, when $\gamma = 2.00$, the intensive margin accounts for over 60% of the total decrease in hours, while when $\gamma = .10$ this value is less than 5%. If a researcher used the benchmark model from Section 12.2 to interpret steady state differences in aggregate hours worked across two Rogerson-Wallenius economies with the same low value of γ but differing scales of their tax and transfer system, they would infer a value of γ that is more than an order of magnitude larger than the true underlying value of γ. This happens because the change in aggregate hours includes responses on both the extensive and intensive margins and the implied value of γ must proxy for adjustment along both margins. The greater is the adjustment along the extensive margin, the greater the disconnect between the true value of γ and the value of γ that the intensive margin only model will infer from the aggregate data. A key message from the analysis of Chang and Kim (2006) continues to hold in this setting that features adjustment along both intensive and extensive margins: there is a large disconnect between the parameters that characterize aggregate and individual labor supply. In particular, labor supply elasticities estimated on micro panel data using workers with positive hours are not particularly informative for predicting the aggregate effects of permanent changes in taxes. Moreover, the aggregate elasticity is large.

Because the Rogerson and Wallenius (2009) model is somewhat stylized, it is of interest to consider robustness to allowing for richer and more realistic empirical specifications. Erosa, Fuster, and Kambourov (2016) go quite far in assessing this. Specifically, they extend the Rogerson-Wallenius model along many dimensions in order to match a wide variety of features of wages and hours worked for males between the ages of 25 and 61. Their analysis allows for multiple sources of heterogeneity (both idiosyncratic shocks as in Chang and Kim (2006) and permanent productivity differences), multiple nonconvexities (fixed utility costs of working in addition to nonconvex earnings), time aggregation, and measurement error in wages. While these features do matter for the empirical properties of the model and its ability to replicate the salient features of the data, the conclusions are broadly similar. They find that the aggregate labor supply elasticity to a temporary unanticipated wage change is 1.75.

Chapter 13

Growth

Timo Boppart and Peter J. Klenow

13.1 Motivation

In this chapter we will study long-run economic growth. Figure 13.1, which replicates Figure 2.1 introduced in Chapter 2, illustrates the striking phenomenon of sustained growth in U.S. GDP per capita over the past two centuries at a steady rate of about 1.5 percent a year. What is behind this pattern of steady growth?

Standards of living have improved not only in advanced economies like the U.S., but also in developing economies in the last half-century. Countries did not, however, all grow at the same rate or exhibit the same income levels. The chapter also deals with the determinants of differences in GDP per capita levels around the globe.

In this chapter we first take a look at some stylized facts about growth and development in the U.S. and elsewhere. We then lay out a version of the neoclassical growth model with investment-specific technical change and use this framework to analyze the data. Finally, we move on to endogenous growth theory, discussing canonical models and how they can be brought to the data.

Motivating the study of economic growth is straightforward. The process of economic growth transformed the U.S. far beyond what is visible in the economic indicators. Understanding the forces behind this phenomenon is of first-order importance to human welfare. Moreover, understanding why some countries are poor and some countries are rich and what policy could do to influence this is a many-trillion dollar question for the entire social sciences. A nice quote that makes this point is from Lucas Jr (1988): "Is there some action a government of India could take that would lead the Indian economy to grow like Indonesia's or Egypt's? If so, what, exactly? If not, what is it about the 'nature of India' that makes it so? The consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else."

13.2 Empirical patterns

As mentioned, Figure 13.1 depicts the steady growth of U.S. GDP per capita over the last two centuries. Appendix Figure 13.A.1 makes it clear that one should not take such growth

for granted, as it globally began slowly around a thousand years ago before accelerating over the last 200 years (after the so called Industrial Revolution).

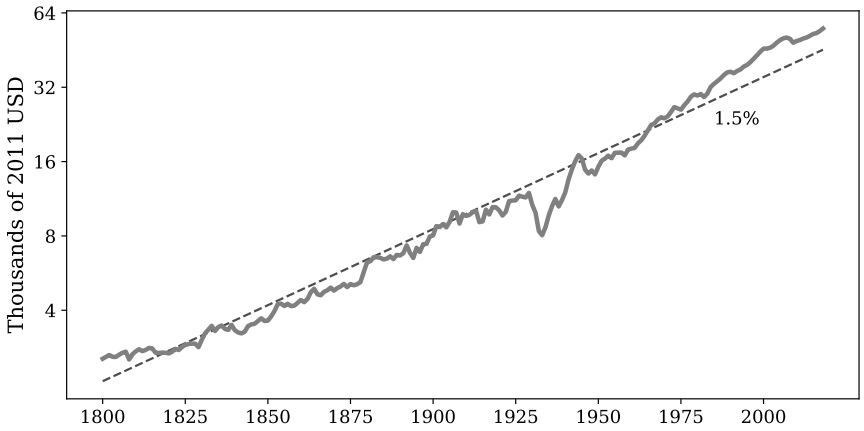


Figure 13.1: U.S. Real GDP per Capita

Source: Maddison Project Database (2020). The vertical axis is in ratio scale.

We can get a sense of whether growth was faster or slower in other countries compared to the U.S. using data from the Penn World Table. Figure 13.2 plots GDP per worker for regions of the world *relative to the U.S.* from 1960 through 2019. The U.S. is normalized to 1 in each year. Each region has a balanced panel of countries, with the countries weighted by their employment. One can see that Europe converged toward the U.S. from 1960 to 1990 or so, but since has grown parallel to the U.S. Latin America as a whole mostly moved sideways over the sample. East Asia, led by China, Japan and South Korea, rose from 1/16th of U.S. GDP per worker in 1960 to almost 1/4th in 2019. Sub-Saharan Africa grew at a similar rate to the U.S. from 1960 to 1980, then fell behind from 1980 to 2000 (falling from 1/8th to 1/16th) before stabilizing. Finally, South Asia tracked at about 1/16th of the U.S. until the mid-1970s, lost ground through the mid-1980s, then soared to over 1/8th led by India's surging growth. The broad picture that emerges is that most regions generally grew along with the U.S., though at differing rates.

Figure 13.3 provides a more detailed country-by-country look. The vertical axis is GDP per capita in 2019 and the horizontal axis is GDP per capita in 1960, with the U.S. normalized to 1 in both years. Countries along the 45 degree line maintained the same growth rate as the U.S. This was the "typical" pattern in that a regression of log 2019 GDP per capita on the 1960 version yields a coefficient of about 1. That said, many countries grew considerably faster (e.g., South Korea or Singapore) or slower (e.g., the Congo or Burundi). Rich countries generally hewed close to the U.S. growth rate, whereas developing countries exhibited more dispersion.

Figure 13.3 displays neither strong divergence nor strong convergence of per capita incomes over time. The early empirical cross-country growth literature emphasized a distinction between unconditional and conditional convergence. As we already saw in Figure 3.6 earlier, this figure suggests no unconditional convergence.[1] The absence of divergence is

[1]The literature further distinguished "beta" and "sigma" convergence. Beta convergence refers to a

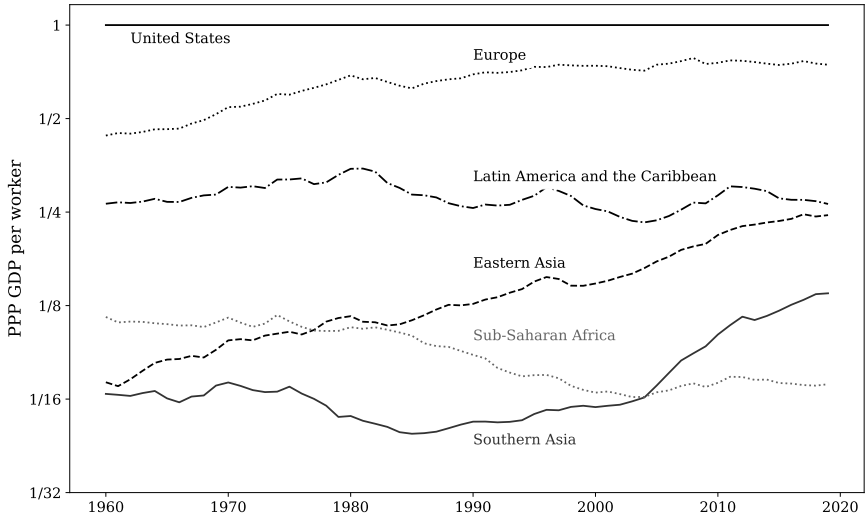


Figure 13.2: PPP GDP per Worker (U.S. = 1)

Source: Penn World Table 10.0. PPP GDP per worker is calculated as the ratio of the "rgdpo" and "emp" variables relative to the U.S. in all years.

consistent with a common trend, say due to technology that gradually diffuses across countries. See Nath, Ramey, and Klenow (2023) for evidence that country income differences are persistent, whereas country growth differences are largely transitory.

Rather than looking at each country, it is useful to plot the distribution of income across the world population. Appendix Figure 13.A.2 does so under the simplifying assumption that income is the same within a country. The U.S. is normalized to 1 in each year. One can see the distribution shifting to the right because of the rapid growth in China and India. The distribution is left-skewed but decreasingly so over time.[2]

The plots with individual countries highlight the tremendous dispersion in living standards across countries. In Appendix Figure 13.A.4 we focus on the PPP GDP per capita in the 20 largest countries by population in 2019. With the U.S. normalized to 1, one can see that income differs by a factor of 64 between the Congo and the U.S. Even with the rapid growth in China and India in recent decades, they have only attained about 1/4th and 1/8th the level of U.S. per capita income in 2019.

These facts on growth rates and income differences call out for explanation. In a deeper sense they may stem from policies such as taxation; government investments in education, health, and infrastructure; regulation and government ownership; and openness to trade and foreign direct investment. But short of that, we can do growth and development accounting to determine the more proximate sources of growth and development differences.

To be concrete, imagine a simple Cobb-Douglas aggregate production function that is

tendency for countries with initially higher per capita incomes to grow more slowly. Sigma convergence involves a narrowing of the dispersion of per capita incomes over time. One could have beta convergence without sigma convergence because of ongoing shocks. But our Figure 13.3 and the literature find a lack of both beta and sigma convergence.

[2]For contrast, Appendix Figure 13.A.3 plots the histogram in the case when each country is weighted equally. The distribution still shifts to the right, but is much less skewed in each year.

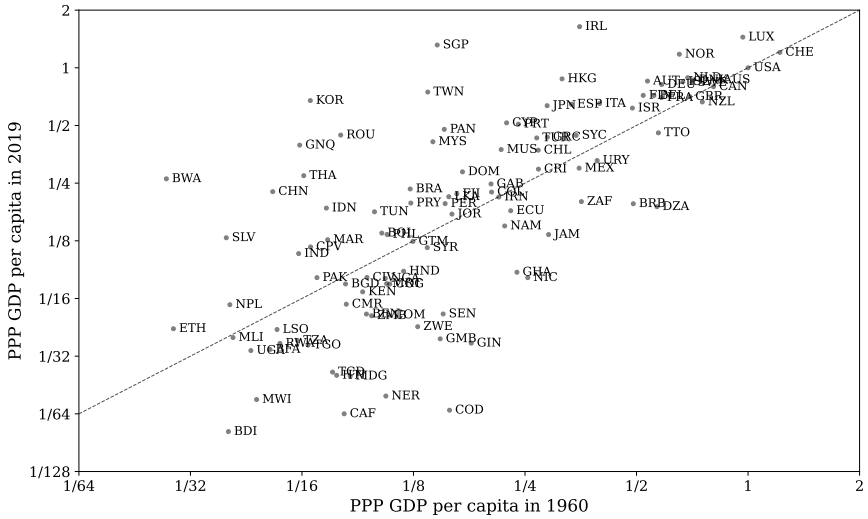


Figure 13.3: PPP GDP per Capita in 1960 and 2019 (U.S. = 1)

Note: The data comes from the Penn World Table 10.0. PPP GDP per capita is calculated as the ratio of the "rgdpo" and "pop" variables relative to the U.S. in both years. The dotted line is the 45 degree line. The simple correlation across the two years is 0.75 and the elasticity is 0.97 (0.091).

common to all countries i and years t:

$$Y_{it} = K_{it}^{\alpha} \left(A_{it} H_{it}\right)^{1-\alpha}, \tag{13.1}$$

where Y is real output, K is real physical capital, H is total human capital (efficiency units summed across all workers), and A is residual labor-augmenting TFP. Here α is the production elasticity of output with respect to physical capital.[3]

Dividing equation (13.1) by L and rearranging implies

$$\frac{Y_{it}}{L_{it}} = \left(\frac{K_{it}}{Y_{it}}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{H_{it}}{L_{it}}\right) A_{it}. \tag{13.2}$$

Here L is total employment so that H/L is (average) human capital per worker. As described in Section 4.3.3 above, the balanced growth path of a neoclassical growth model features a stable K/Y that does not depend on the levels of A or H/L. Meanwhile, models such as Mankiw, Romer, and Weil (1992) yield a stationary level of human capital per worker that does not depend on A or K/Y. Hence, the decomposition in (13.2) is compatible with steady states under standard forms of endogenous investments in physical and human capital. In the data, there is substantial persistence in capital-output ratios over time within countries (see Appendix Figure 13.A.5). This is consistent with countries being close to their respective balanced growth paths in terms of capital accumulation. Countries do not seem to transition

[3]Assuming a Cobb-Douglas functional form is convenient but not strictly necessary. As in Solow (1957), one can calculate residual TFP for any neoclassical (i.e., constant returns) production function. Moreover, even with a Cobb-Douglas specification the production elasticity could vary over time or across countries.

to a common capital-output ratio. For this reason we will do the accounting below based on (13.2) rather than (13.1).[4]

Taking logs and differentiating with respect to time we obtain a form convenient for growth accounting:

$$\Delta \log\left(\frac{Y_{i,t}}{L_{i,t}}\right) = \frac{\alpha}{1-\alpha}\Delta \log\left(\frac{K_{i,t}}{Y_{i,t}}\right) + \Delta \log\left(\frac{H_{i,t}}{L_{i,t}}\right) + \Delta \log\left(A_{i,t}\right). \tag{13.3}$$

One can then average the right side components over time, back out TFP growth $\Delta \log(A)$, and compare it to the left hand side $\Delta \log(Y/L)$. Table 13.1 does this for the U.S. in recent decades. The table applies to total private businesses for the years 1948–2020 and α is approximated by physical capital's cost share.[5] Increases in the capital-output ratio contributed modestly—less than one-tenth of average growth over the entire sample. Human capital per worker contributed twice as much, primarily through rising years of education. Still, this leaves about three-fourths of growth coming from residual TFP. This echoes what Solow (1957) famously found in earlier U.S. data.[6]

Table 13.1 further shows that TFP growth accounts for most of the medium-run shifts in the rate of growth in output per hour over time in the U.S. When growth is high (such as from 1948–1973 or 1995–2007) or low (1973–1995 or 2007–2020), it is primarily due to the pace of TFP growth. Researchers such as Jorgenson, Ho, and Stiroh (2008) have attributed TFP growth from 1995–2007 to the contribution of ICT (Information and Communication Technology), both directly in the ICT-producing sector and downstream in ICT-using sectors such as retail trade.[7]

By taking logs of equation (13.2) we can decompose differences in the level of development across countries at a point in time. In doing so the Penn World Table follows Klenow and Rodriguez-Clare (1997), Hall and Jones (1999), and Bils and Klenow (2000) in using an efficiency units formulation (all levels of human capital are perfectly substitutable) for simplicity. Their method, unlike that of Mankiw et al. (1992), explicitly ties human capital to evidence on the Mincerian wage return to years of schooling.

Table 13.2 looks across 117 countries in 2019. This exercise is dubbed income or development accounting as it backs out TFPs as in Solow's growth accounting method but is applied to levels as opposed to growth rates. The columns pertain to log GDP per worker, the log capital-output ratio multiplied by (1−labor's share)/(labor's share), the log of human capital per worker, and residual labor-augmenting TFP.[8] The second row reports the elasticity of each log variable with respect to GDP per worker. These elasticities add up

[4]Doing accounting based on (13.2) attributes a larger role to differences in TFP and human capital relative to physical capital. Equation (13.2) takes into account that higher levels of TFP and human capital will induce more physical capital accumulation in steady state.

[5]Using a cost share as opposed to an income share prevents any price-cost markups that affect revenue relative to costs from biasing our estimate of the elasticity.

[6]Because of its mysterious nature, Abramovitz (1956) dubbed TFP a "measure of our ignorance." Below we try to unpack some of what we have learned about determinants of TFP.

[7]This resolved the Solow Paradox that "you can see the computer age everywhere but in the productivity statistics." See Solow (1987).

[8]Under competitive markets we can approximate the production elasticity of output with respect to labor using payments to labor relative to GDP. We average this across countries to arrive at a single estimate.

Table 13.1: Growth Accounting for the U.S.

		Contributions from		
Period	Y/L	K/Y	H/L	A
1948–2020	2.37	0.21	0.38	1.79
1948–1973	3.28	-0.18	0.27	3.19
1973–1995	1.54	0.46	0.36	0.72
1995–2007	2.80	0.32	0.40	2.08
2007–2020	1.64	0.43	0.59	0.63

Note: The data comes from the U.S. Bureau of Labor Statistics. Y/L denotes real output per hour, K/Y the real capital-output ratio, H/L human capital per worker (which grows predominantly from rising years of schooling), and A is residual labor-augmenting TFP inclusive of contributions from R&D and intellectual property. The contribution of physical capital is scaled by $\alpha/(1-\alpha)$, where α is the average cost share for physical capital over the sample, equal to 0.34. See equation (13.2).

to 1 by construction. The coefficients are the same as the contribution of each term in a variance decomposition in which we split the covariance terms between all three components equally. By this metric, capital intensity generates only 14% of differences in income across countries. Schooling is more important, being responsible for 22% of differences. Like the U.S. over time, however, the largest contribution is from residual TFP at 64%.

The last row of Table 13.2 gives the 90/10 ratios of each variable, where 90 and 10 refer to the 90th and 10th percentile countries in terms of 2019 PPP GDP per worker. An advantage of looking at 90/10 ratios is that they are less sensitive to outliers. Remarkably, if we put the ratios in log space, they imply almost the same contributions. For example, residual TFP contributes around 64% ($\log(4.92)/\log(12) \approx 0.64$).

Studies incorporating differences in the quality of schooling (Schoellman, 2012) and human capital accumulated on the job (Lagakos, Moll, Porzio, Qian, and Schoellman, 2018) arrive at larger contributions from human capital—on the order of 50%—thereby winnowing the role of residual TFP down to about 40%. Appendix Figure 13.A.7 provides the elasticities in Table 13.2 by year. They are broadly similar in 1960 and 2019, though the importance of human capital has diminished whereas the role of capital intensity fell from 1990 to 2010 before rising.

13.3 Neoclassical growth with investment-specific technical change

Chapter 2 discussed long-run stylized facts observed in advanced economies (sometimes referred to as the Kaldor facts) and how they discipline the typically imposed structure in growth models. The Solow model in Chapter 3, for example, settles down to a stable capital-output ratio. The neoclassical growth model, as developed by David Cass and Tjalling Koopmans in the 1960s, microfounded the consumption-saving decision and estab-

Table 13.2: Development Accounting in 2019

		Contributions from		
Statistic	Y/L	K/Y	H/L	A
Variance of log	1.00	0.14	0.08	0.57
Elasticity wrt Y/L		0.14	0.22	0.64
90/10 ratio	12.00	1.40	1.74	4.92

Note: The data comes from the Penn World Table 10.0. The sample is 117 countries in 2019. Output per worker is constructed using the "rgdpo" and "emp" variables. The capital to output ratio is constructed using the "cn", "rgdpo" and "labsh" variables. The human capital index corresponds to the "hc" variable. We use the population-weighted average labor share of 0.53 across countries in 2019, which implies $\alpha = 0.47$.

lished conditions for a steady state optimal saving rate.[9] As the aggregate saving rate is not equal to a universal constant but rather determined by the behavior of many agents in an economy, micro-founding the corresponding intertemporal decision is important in the light of a classic Lucas critique (see Chapter 1). Furthermore, specifying intertemporal preferences allows us to use the theory to make normative statements about welfare.

In the following we specify and solve a specific version of the neoclassical growth model in continuous time that allows for investment-specific technical change (as in Greenwood, Hercowitz, and Krusell, 1997). This framework is then used to address the time series and cross-country data. We choose a version of the model with investment-specific technical change because it speaks to a striking pattern in the data on relative prices between the two distinct final uses: consumption and investment.[10] In the U.S. the relative price of investment fell systematically since 1970 at an average annual rate of about 1.3% (see Figure 13.4). Investment can be further split up into equipment and structures. The decrease in the relative price of investment is driven by equipment which saw a decrease of its relative price to consumption at almost 2.5% a year. In contrast, the relative price of structures (the left out category in Figure 13.4) increased over time. Figure 13.5 shows that the relative prices also systematically differ across countries. In the poorest countries the price of investment relative to consumption is roughly twice as high as in the U.S. This suggests that the poorest countries are particularly inefficient in turning primary production factors into investment goods (Hsieh and Klenow, 2007). This has consequences for development accounting as we expect richer countries with the same nominal saving rate to arrive on average at a higher real capital-output ratio. The model we lay out in this section allows us to speak to these facts.

Preferences As in Chapter 9, we denote a variable X that changes over time as $X(t)$ and its time derivative $\dot{X}(t) \equiv dX(t)/dt$. We consider a representative household, and abstract

[9]The intertemporal problem of optimal savings had actually been studied much earlier by Ramsey (1928), far before Solow wrote his seminal paper. The neoclassical growth model is therefore often called the Ramsey-Cass-Koopmans model, or just the Ramsey growth model.

[10]The plain vanilla version of the neoclassical growth model is discussed in Chapter 4.

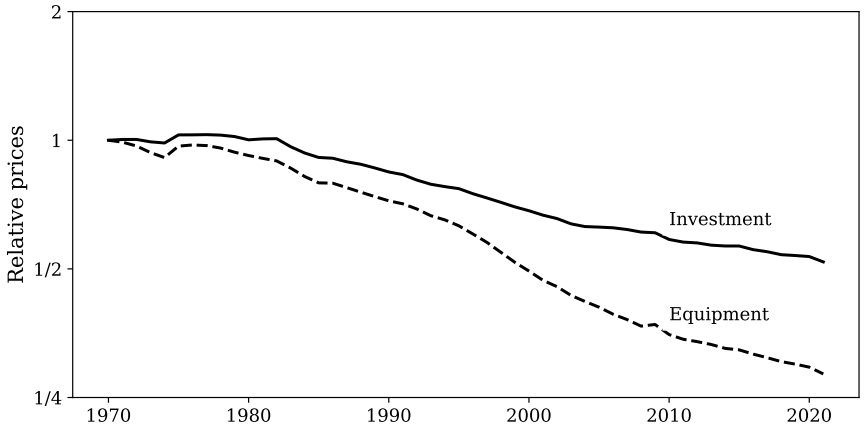


Figure 13.4: Relative price of equipment and investment in the U.S.

Notes: The data comes from the U.S. Bureau of Economic Analysis (NIPA Table 1.1.4). The series are relative to the price of consumption, and are each normalized to 1 in 1970. The price level for structures nearly doubled over 50 years by growing 1.2% per year relative to the consumption deflator. In contrast, the annual growth rate of the relative price of equipment and total investment were about -2.5% and -1.3%, respectively.

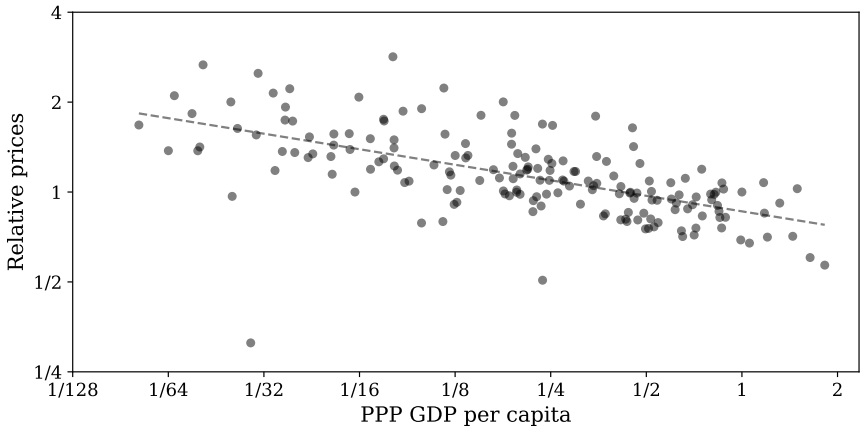


Figure 13.5: Price of investment relative to consumption in 2019 across countries

Notes: The data comes from the Penn World Table 10.0. The U.S. is normalized to 1 for both variables. The price of investment relative to consumption is calculated from the "pl_i" and "pl_c" variables. PPP GDP per capita is calculated as the ratio of the "rgdpo" and "pop" variables. The estimated elasticity is equal to -0.17 with a standard error of 0.02.

from endogenous labor supply. The household problem is given by

$$\max_{\{a(t),c(t)\}_{t=0}^{\infty}} \int_0^{\infty} e^{-(\rho-n)t} \frac{c(t)^{1-\sigma}-1}{1-\sigma} dt, \tag{13.4}$$

subject to

$$\dot{a}(t) = (r(t) - n)a(t) + w(t)h - c(t), \forall t, \text{ with } a(0) \text{ given},$$

and the no-Ponzi game condition

$$\lim_{T\to\infty} \left\{ e^{-\int_0^T (r(\tau)-n)d\tau} a(T) \right\} \geq 0.$$

Here ρ is the discount rate and n is the population growth rate. The initial population size is normalized to unity. The period budget constraints are stated in per-capita terms so a denotes per-capita wealth and h the level of human capital per worker. The household's problem takes the form that is familiar from Chapter 9.

In the household budget the price of the consumption good is normalized to one. The household converts its savings in investment goods at price P_x, and rent its capital stock out to firms. The household receives from the firm the rental rate $R(t)$ but has to incur depreciation δ on its capital plus faces a change in the relative price of $\dot{P}_x(t)$. Hence, the condition that links the return r to the rental rate, the depreciation rate, and the investment price reads

$$P_x(t)r(t) = R(t) - \delta P_x(t) + \dot{P}_x(t). \tag{13.5}$$

Technology Final output Y is competitively produced by a representative firm with the Cobb-Douglas technology

$$Y(t) = K(t)^{\alpha} \left(A_y e^{\gamma_y t} L(t) \right)^{1-\alpha}. \tag{13.6}$$

Here K denotes capital input and L labor input in production. The parameter $\alpha \in (0, 1)$ controls the output elasticity of capital. The term $A_y e^{\gamma_y t}$ captures the technology in Harrod-neutral form (consisting of a level A_y and a rate of technological change $\gamma_y > 0$).

Output Y can be transformed one for one into consumption goods and one for $A_x e^{\gamma_x t}$ into investment goods. As both technologies are linear in Y goods, the price of Y will equalize to the price of C under perfect competition, which we choose as a numéraire, i.e., $P_y(t) = P_c(t) = 1, \forall t$ and the price of the investment good will be given by $P_x(t) = e^{-\gamma_x t}/A_x$. The object Y is then a measure of GDP expressed in terms of consumption units. Here $\gamma_x \geq 0$ represents investment-specific technical change; with $\gamma_x > 0$ the relative price of investment declines over time. The production side can then be characterized as a static profit maximization problem of a representative firm:

$$\pi(t) = \max_{K(t),L(t)} \left\{ K(t)^{\alpha} \left(A_y e^{\gamma_y t} L(t) \right)^{1-\alpha} - R(t)K(t) - w(t)L(t) \right\}. \tag{13.7}$$

Note that, given our choice of numéraire, (per-capita) wealth a and all the prices w, r, R, and P_x are stated in relative terms and measured in units of consumption goods.

Equilibrium definition A competitive equilibrium is then defined as a path of prices and quantities that:

1. solves the household problem (13.4) and the firm problem (13.7).
2. fulfills asset and labor markets clearing conditions, i.e., $K(t)e^{-\gamma_x t}/A_x = a(t)e^{nt}$ and $L(t) = e^{nt}h$.

3. fulfills the return condition $r(t) = R(t)A_x e^{\gamma_x t} - \delta - \gamma_x$.

Here $a(t)e^{nt}$ is aggregate wealth (in units of the consumption good). This amount is invested in capital goods at price $e^{-\gamma_x t}/A_x$. Finally, to arrive at the return condition in (3.) we replaced $P_x(t)$ in (13.5) by $e^{-\gamma_x t}/A_x$.

As the model features growth we need to impose the following restriction to ensure that the household problem is well-defined[11]

$$n - \rho + \frac{\alpha(1-\sigma)}{1-\alpha}\gamma_x + (1-\sigma)\gamma_y < 0. \tag{13.8}$$

Model solution The household's problem can be solved using the technique we learned in Chapter 9. In Appendix Section 13.A.1 we solve for the equilibrium dynamics and show that they boil down the following differential equations in $K(t)$ and $c(t)$:

$$\frac{\dot{c}(t)}{c(t)} = \frac{A_x e^{\gamma_x t}\alpha\left(\frac{A_y e^{(\gamma_y+n)t}h}{K(t)}\right)^{1-\alpha} - \delta - \gamma_x - \rho}{\sigma} \tag{13.9}$$

and

$$\frac{\dot{K}(t)}{K(t)} = A_x\left(\left(\frac{e^{(\gamma_x/(1-\alpha)+\gamma_y+n)t}A_y h}{K(t)}\right)^{1-\alpha} - \frac{e^{(\gamma_x+n)t}c(t)}{K(t)}\right) - \delta. \tag{13.10}$$

The initial capital stock $K(0)$ is exogenously given and the additional terminal condition is given by

$$\lim_{T\to\infty}\left\{\frac{K(T)}{A_x}e^{-(\gamma_x+\rho)T}c(T)^{-\sigma}\right\} = 0. \tag{13.11}$$

Equation (13.9) is the consumption Euler equation whereas (13.10) states the law of motion of the capital stock. As the welfare theorems apply in this framework the same system of differential equations in $K(t)$ and $c(t)$ could have been found more directly by solving the planner's problem.[12] However, as we plan to confront the theory's prediction with data on prices and real quantities it is essential that we have solved for the decentralized equilibrium.

Balanced growth path We define a balanced growth path the standard way as a path along which all prices and quantities grow at constant rates. Does this economy admit a balanced growth path and if so at what rates do physical capital K, output Y and per-capita consumption c grow? Given that the model features investment-specific technical change these questions are non-trivial. Can the right-hand side of (13.10) be constant for all t? One notes that a constant ratio $e^{(\gamma_y+n)t}/K(t)$ is not consistent with a constant growth rate in K. Hence, the standard candidate in which capital grows at the combined rate of population and Harrod-neutral technological change is inconsistent with a balanced growth path. However, a valid candidate is with

$$\frac{\dot{K}(t)}{K(t)} = \gamma_x/(1-\alpha) + \gamma_y + n \equiv g_K$$

[11] We will see below that the long-run growth rate in per-capita consumption is $\frac{\alpha}{1-\alpha}\gamma_x + \gamma_y$. Then, condition (13.8) ensures that utility is bounded.

[12] Appendix Section 13.A.2 states and solves the planner problem for this economy.

and

$$\frac{\dot{c}(t)}{c(t)} = \alpha\gamma_x/(1-\alpha) + \gamma_y \equiv g_c.$$

As a consequence, we can write the system of differential equations in detrended variables $\tilde{c}(t) \equiv c(t)/e^{g_c t}$ and $\tilde{k}(t) \equiv K(t)/e^{g_K t}$ as:

$$\frac{\dot{\tilde{c}}(t)}{\tilde{c}(t)} = \frac{\alpha A_x \left(\frac{A_y h}{\tilde{k}(t)}\right)^{1-\alpha} - \delta - \gamma_x - \rho}{\sigma} - \frac{\alpha\gamma_x}{1-\alpha} - \gamma_y, \tag{13.12}$$

$$\frac{\dot{\tilde{k}}(t)}{\tilde{k}(t)} = A_x \left(\left(\frac{A_y h}{\tilde{k}(t)}\right)^{1-\alpha} - \frac{\tilde{c}(t)}{\tilde{k}(t)} \right) - \delta - \frac{\gamma_x}{1-\alpha} - \gamma_y - n, \tag{13.13}$$

with $\tilde{k}(0)$ given and the terminal condition[13]

$$\lim_{T\to\infty} \left\{ \tilde{k}(T)\tilde{c}(T)^{-\sigma} e^{(n+(1-\sigma)\alpha\gamma_x/(1-\alpha)+(1-\sigma)\gamma_y-\rho)T} \right\} = 0.$$

This system of differential equations is indeed consistent with a stationary point that we can denote by $\tilde{c}^\star$ and $\tilde{k}^\star$. The terminal condition is fulfilled along this balanced growth path as we imposed condition (13.8). By setting $\dot{\tilde{k}}(t)/\tilde{k}(t)$ and $\dot{\tilde{c}}(t)/\tilde{c}(t)$ equal to zero we get

$$\tilde{k}^\star = A_x^{\frac{1}{1-\alpha}} A_y h \left(\frac{\alpha}{\frac{\alpha\sigma\gamma_x}{1-\alpha} + \gamma_x + \sigma\gamma_y + \delta + \rho} \right)^{\frac{1}{1-\alpha}}, \tag{13.14}$$

and $\tilde{c}^\star$ follows directly from

$$\tilde{c}^\star = \left(\tilde{k}^\star\right)^\alpha (A_y h)^{1-\alpha} - (g_K + \delta)\frac{\tilde{k}^\star}{A_x}. \tag{13.15}$$

In general, in a neoclassical growth framework with CRRA preferences any production function which fulfills standard assumptions can support a balanced growth path. However, as the model here features investment-specific technical change (with $\gamma_x > 0$) this is no longer the case and the assumption of a Cobb-Douglas technology in (13.6) is key in order to generate a balanced growth path. This relates back to Uzawa's theorem (see Appendix 3.A) that requires all the technological change to be of the labor augmenting type. In the special case of a Cobb-Douglas production function factor augmenting technical change is not distinctly defined and capital-augmenting technical change can as well be expressed in labor-augmenting terms by raising it to the power $\alpha/(1-\alpha)$.

[13] Adding the constant $A_x^{\frac{1}{1-\alpha}} A_y h$ to the denominator of the definition of the detrended capital stock would allow us to write the expression in a more compact way. However, further below we entertain a thought experiment where we allow countries to differ in these constants and it is then helpful to explicitly see how these factors affect the detrended capital stock along the balanced growth path.

BGP predictions Along the balanced growth path a constant fraction $s^\star = \alpha(g_K + \delta)/(\sigma g_c + \gamma_x + \delta + \rho)$ is saved out of income before depreciation.[14] The Solow model simply imposes such a constant saving rate. Here, in contrast, we derive this $s^\star$ endogenously and it will in general move along the transition.

We can compare the steady state capital stock $\tilde{k}^\star$ with the Golden Rule capital stock. Maximizing (13.15) with respect to $\tilde{k}^\star$ gives the Golden Rule capital stock

$$\tilde{k}^{gold} = A_x^{\frac{1}{1-\alpha}} A_y h \left(\frac{\alpha}{g_K + \delta} \right)^{\frac{1}{1-\alpha}}. \tag{13.16}$$

This would be achieved if the entire capital income was saved, i.e., $s^{gold} = \alpha$. As $\frac{\alpha\sigma\gamma_x}{1-\alpha} + \gamma_x + \sigma\gamma_y + \delta + \rho > g_K + \delta$ is ensured by condition (13.8) the steady state capital stock is strictly below the golden rule capital stock, which is reminiscent of the discussion in Appendix 4.A.4.

What are the theory's predictions for key aggregates and prices along the balanced growth path? As the model features two final uses—consumption and investment—and their relative price is changing over time, one has to be careful when evaluating the model prediction for real quantities. Total GDP when measured in units of consumption goods is captured by Y. Similarly, w and r denote the wage and interest rate measured in consumption units.[15] The wage rate grows at the same rate as per-capita consumption g_c, whereas the interest rate, $r^\star = \sigma g_c + \rho$ is constant along the balanced growth path (expressing how much consumption is obtained tomorrow by forgiving one unit of consumption today). In contrast, R denotes the rental rate of one unit of physical capital (also measured in terms of tomorrow's consumption units) and is falling along the balanced growth path at constant rate γ_x. The difference is precisely explained by the relative price of investment which falls at rate γ_x too, i.e., investing one unit of physical capital is getting in terms of consumption units less and less costly over time in terms of consumption units.

Also the price of the capital stock relative to the deflator of the consumption good, $e^{-\gamma_x t} A_x^{-1}$ decreases over time at the rate of investment-specific technical change γ_x. The *nominal* capital-output ratio $e^{-\gamma_x t} A_x^{-1} K(t)/Y(t)$ is constant along the balanced growth path and equal to $s^\star/(g_K + \delta)$.[16] "Total real output" is not a traded commodity in the model so a price deflator for total GDP has not been defined in the theory. We can still mimic in the model what statistical offices do in practice: As the nominal shares of investment and consumption are constant along the balanced growth path a reasonable definition of the growth rate of the GDP deflator is $(\dot{P}_x/P_x)^{s^\star}(\dot{P}_c/P_c)^{1-s^\star} = -(1-s^\star)\gamma_x$.[17] With such a definition, it is straightforward to see that the theory predicts that the ratio of real capital to real output grows at rate $(1-s^\star)\gamma_x$.

Transitional dynamics The transitional dynamics are similar as in a version of the neo-calassical growth model without investment-specific technical change. If the economy starts

[14] To see this equate the right-hand side of (13.15) to $(1-s^\star)\left(\tilde{k}^\star\right)^\alpha (A_y h)^{1-\alpha}$ and solve for $s^\star$.

[15] The expressions for the equilibrium wage and interest rate can be found in Appendix 13.A.1.

[16] A similar formula was obtained in the Solow model (see Section 3.2). Importantly, here the formula applies to the nominal capital-output ratio and not to a measure of the real capital-output ratio.

[17] The second expression follows as we chose P_c as numéraire.

out with a detrended capital stock below $\tilde{k}^\star$ capital accumulates faster and $\tilde{k}$ increases over time until it asymptotically reaches $\tilde{k}^\star$.

For the special case with $\sigma = \alpha$ the model allows for a closed-form solution along the transition. This be shown by guessing and verifying the system (13.12) and (13.13) is globally consistent with the consumption rule $\tilde{c}(t) = \tilde{k}(t) \cdot (\delta(1-\alpha) + \gamma_x(1-\alpha) - \alpha n + \rho) / (\alpha A_x)$. For more general cases where no closed-form solution exists one can still analytically analyze the local dynamics by linearizing the system around the steady state. This is illustrated in Appendix 13.A.3. The linearized system can also be used to assess the (local) speed of convergence. A simple calibration done in Appendix 13.A.3 suggests that the speed of convergence is fast, implying a half-life of about 5.5 years.

Discussion of the framework The model we laid out assumes a closed economy. One obvious justification of this assumption is that the world as a whole is indeed a closed economy. For large economies like the U.S. international trade flows are quantitatively not that large relative to GDP and trade may be considered not to be of first-order importance for investment and capital accumulation. It is a standard approach to address the cross-country data through the lens of closed economy theory and this is also what we will pursue below.[18]

The framework we laid out assumes a representative household but as the CRRA utility function implies overall homothetic utility, preferences allow for aggregation.[19] The CRRA period utility function is key in order to microfound a constant long-run saving rate (with growth). We abstracted from endogenous labor supply. However, as discussed in Section 2.1.8, an endogenous labor supply decision could be allowed for and can be reconciled with a balanced growth path (see Boppart and Krusell, 2020).

The level of human capital h is assumed to be constant but the model allows to study transitional movements in this level. Allowing instead for sustained growth in h would potentially lead to very different dynamics which we comment on in the section below on endogenous growth.

The model consists of multiple sectors (an investment sector and a consumption sector). However, along the balanced growth path the nominal investment rate is constant which precludes a systematic long-run reallocation of production factors across sectors—a phenomenon called structural change. It is indeed a salient empirical observation that development is accompanied by shifts in employment shares, first from agriculture to manufacturing and at a later stage into services. How to square this unbalanced feature of growth at the sectoral level with the balanced nature of growth in the aggregate? There indeed exist restrictions on technologies and preferences such that there exists a balanced growth path at the aggregate level despite structural change at the sectoral level (see Kongsamut, Rebelo, and Xie, 2001, Ngai and Pissarides, 2007, and Boppart, 2014).[20]

[18] The interaction between countries is arguably more important to influence "technology". We will comment on technological spill-overs and adoption further below when we discuss theories of endogenous growth.

[19] I.e., we could as well assume that there is a unit interval of heterogeneous households endowed with h_i units of human capital and $a_i(0)$ units of initial wealth. As long as $\int_0^1 h_i \, di = h$ and $\int_0^1 a_i(0) \, di = a(0)$ the aggregate dynamics would be identical to the ones in the economy specified above.

[20] The standard definition of an exact balanced growth path as in Section 3.2 needs to be relaxed when studying structural change as imposing that all aggregate variables have to grow at constant rates restricts

Finally, a relevant extension is to explicitly modeling energy and natural resources as production factors. Chapter 25 discusses such models.

13.4 Confronting the investment-specific technical change model with the data

How does the model in the previous section allow us to think about the data? What explains the big income differences across countries? What is behind the observed growth miracles and growth disasters? By addressing these questions we also comment on the extensive empirical literature on growth.

One way to think about income differences through the lens of the theory is as transitional dynamics due to differences in the *initial* capital stock. Suppose all countries are identical except for their initial capital stock. As all countries then converge to the same balanced growth path, we expect to see a strong pattern of convergence of income levels across countries. As emphasized earlier, however, we do not find strong support for absolute (i.e., unconditional) convergence in the cross-country data (Figure 13.3). And even in samples where there is evidence of unconditional convergence (such as across U.S. states from 1880 to 1980, among OECD countries from 1950 to 1980, or in the full cross-country sample since the 2000s), the observed speed of convergence is not nearly as fast as the one implied by the theory.[21]

At the same time the world log income distribution does not systematically fan out over time (see Appendix Figure 13.A.2). Hence, there are clear forces that hold the income distribution together over time (and convergence due to the diminishing marginal product of capital is one such force). The lack of *divergence* in Figure 13.3 suggests the rate of technical change does not differ systematically between rich and poor countries. It is therefore reasonable to think of all countries to share the same parameters γ_y and γ_x in the model above. The labor income share is across countries not systematically related to the level of development suggesting that the parameter α can also be treated as common across countries. Technology differences could still matter, but would rather be reflected in the model by *level* differences in A_y or A_x.

Why do countries persistently differ in their income levels? One possibility is capital-output ratios, say, e.g., due to country differences in population growth n.[22] Moreover, given common values of γ_y, γ_x, and α, different levels of human capital h, A_y, or A_x lead to different

the sectoral shares to be constant. One such modification of the definition is to define an exact balanced growth path as a path along which the physical capital stock grows at a constant rate (see Alder, Boppart, and Müller, 2022). Alternatively, one could simply not insist on exact balanced growth but rather rely on such a path to exist asymptotically.

[21] Remember, a simple calibration in the model above suggests a half-life of about 5.5 years. Barro and Sala-i Martin (1992) provide evidence of convergence across U.S. states, Baumol (1986) across OECD countries, and Kremer, Willis, and You (2022) in the full cross-country sample over the past 25 years.

[22] See the expression for the capital stock along the balanced growth path in (13.14) to see how the rate of population growth affects the capital-output ratio.

income levels. One can express output per worker (measured in consumption units) as

$$\frac{Y(t)}{e^{nt}} = \left(\frac{P_k(t)K(t)}{Y(t)}\right)^{\frac{\alpha}{1-\alpha}} \cdot h \cdot \underbrace{A_y A_x^{\frac{\alpha}{1-\alpha}}}_{=A} \cdot e^{\left(\gamma_y + \gamma_x \frac{\alpha}{1-\alpha}\right)t}, \tag{13.17}$$

where $P_k(t) = A_x^{-1} e^{-\gamma_x t}$. Along a balanced growth path, the nominal capital-output ratio $\frac{P_k(t)K(t)}{Y(t)}$ is constant and equal to $s^\star/(g_K+\delta) = \alpha/\left(\frac{\alpha\sigma\gamma_x}{1-\alpha} + \gamma_x + \sigma\gamma_y + \delta + \rho\right)$ and independent of the levels of human capital or TFP. For a given α, equation (13.17) then decomposes per-worker income differences in terms of consumption units in a theory-consistent way into contributions from physical capital, human capital, and a residual technology term. The results are shown in Appendix Table 13.A.1. The message that a majority of income difference is unaccounted for by physical and human capital is reinforced compared to the results in Table 13.2. The share explained by residual TFP is 72% compared to 64% before. This is because the relative price of capital (compared to the GDP deflator) is higher in poorer countries. The contribution of physical capital falls to 7% from 14% previously (whereas the contribution of human capital is unchanged at 22%).

An advantage of this two-sector development accounting is that one can decompose the residual A into general TFP differences A_y and differences specific to the investment sector A_x. One way to discipline the differences in A_x is to bring in the relative prices P_x/P_c from Figure 13.5. The data suggests that the elasticity of $A_x^{\alpha/(1-\alpha)}$ with respect to $P_Y Y/(P_C L)$ is 0.10. This implies that on average 0.10 out of 0.72 or about 14% of the A differences are due to rich countries being particularly efficient at producing investment goods.

Transition dynamics due to capital accumulation might be useful for thinking about the rapid catch-up growth of certain countries. For instance, Japan's GDP per capita relative to the U.S. rose from about 20% in 1950 to 80% in 1990. Rapid capital deepening should, however, be accompanied by a sharp decline in the return to capital, which was not observed in Japan.[23] Thus, even though convergence due to the diminishing marginal product of capital is an important feature of the neoclassical growth model, transitional dynamics due to different initial physical capital levels cannot account for growth miracles such as Japan's. The more realistic way to think about episodes of fast (catch-up) growth is evidently to see them as transitional changes in A_y or A_x. The previous section provided some analytics for such a theory.

Similar to development accounting, (13.17) can be used to decompose growth in per-capita output (in consumption units) in the U.S. over time into contributions from capital deepening, improvements in labor "quality", and residual TFP growth. Appendix Table 13.A.2 presents the results. Over the post-war period the U.S. nominal capital-output ratio was remarkably stable. In contrast, we saw in Table 13.1 that there was modest capital deepening in terms of the U.S. real capital-output ratio (in particular since the 1970s). The contribution of human capital is, again, largely unaffected compared to Table 13.1. This leaves the clear majority of growth accounted for by the TFP residual A. Growth

[23] In order to explain a per-capita income difference of a factor of 5, the relative capital stock per worker needs to fulfill $0.2 = (\tilde{k}_{1950}^{JAP}/\tilde{k}_{1950}^{USA})^{\alpha}$. Everything else equal, based on (13.A.1) the relative rental rate is $(R_{1950}^{JAP}/R_{1950}^{USA}) = (0.2)^{-\frac{1-\alpha}{\alpha}}$. With $\alpha = 1/3$ this implies in 1950 a rental rate in Japan that is 25 (!) times larger than in the U.S. By 1990 this ratio should have declined to less than 2 (see King and Rebelo, 1993).

in per-capita output measured in consumption prices of 2.07% per year can be used to discipline $\gamma_y + \gamma_x \frac{\alpha}{1-\alpha}$ in the model of the previous section. Relative prices allow us to further decompose this growth rate into contributions from γ_y and investment-specific technical change, respectively. Figure 13.4 suggest that $\gamma_x = 0.013$. With an α of about 1/3 this suggests $\gamma_y = 0.014$ and that about 31% of per-capita consumption growth is driven by investment-specific technical change.

To recap, residual TFP accounts for the bulk of differences in income levels and growth rates across countries. What is behind this mysterious residual? For the time series, a natural candidate is technological change (either neutral or investment-specific). In the next section we present theories of endogenous technical change. Technology could also play an important role in income levels across countries. Parente and Prescott (1994), Eaton and Kortum (1996, 1999), and Howitt (2000) are classic models of international technology diffusion. Barro and Sala-i Martin (1995) chapter 8 and Acemoglu (2008) chapter 18 provide textbook treatments. More recent modeling efforts include Akcigit, Ates, and Impullitti (2018), Buera and Oberfield (2020), Hsieh, Klenow, and Shimizu (2022), and Hsieh, Klenow, and Nath (2023).

On the empirical side, Comin and Hobijn (2010) and Comin and Mestieri (2014) provide direct evidence on the use of specific technologies across countries and time. Keller (2004) and Comin and Mestieri (2018) survey the evidence. The Eaton and Kortum papers referenced above document cross-country patenting patterns and how they relate to country size and income. Evenson and Gollin (2003) and Gollin, Hansen, and Wingender (2021) provide evidence for the diffusion of hybrid seeds across many countries, and their impact on agricultural yields. Bloom and Van Reenen (2007) document differences in management practices across countries.

Differences in allocative efficiency could also contribute to income differences across economies. For example, countries may differ in the efficiency with which production factors (capital, labor, intermediates) are allocated across firms due to various market distortions and government policies. Hsieh and Klenow (2009), Alfaro, Charlton, and Kanczuk (2009), and Bartelsman, Haltiwanger, and Scarpetta (2013) provide evidence across countries. See Restuccia and Rogerson (2017) for a survey.

One can ask deeper questions about where these differences in physical capital, human capital, technology, and allocative efficiency come from. A vibrant literature connects these proximate determinants of income to underlying policies and institutions. Acemoglu et al. (2001) is a prominent empirical example, and Chapters 22 and 23 in Acemoglu (2008) survey some of the ways in which institutions and policies can be modeled. Geography is often an underlying factor affecting income directly or indirectly through institutions, such as in Gallup, Sachs, and Mellinger (1999).

13.5 Endogenous growth

AK model The simplest endogenous growth theory is obtained in the economy of the previous section by setting $\alpha = 1$, considering a single output good by setting $A_x = 1$, and shutting down exogenous technological change, i.e., setting $\gamma_y = \gamma_x = 0$. If we premultiply the production function by an additional constant A, it becomes $Y(t) = A \cdot K(t)$, which is

why the literature refers to this as the AK model. The planner's problem of this economy then reads

$$\max_{\{K(t),c(t)\}_{t=0}^{\infty}} \int_0^{\infty} e^{-(\rho-n)t} \frac{c(t)^{1-\sigma}-1}{1-\sigma} dt, \tag{13.18}$$

subject to $\dot{K}(t) \leq AK(t) - e^{nt}c(t) - \delta K(t)$, given $K(0)$, and some non-negativity constraints $K(t) \geq 0, \forall t$. For the problem to be well defined and utility to be bounded we have to assume $\sigma(A-\delta-n) > A-\delta-\rho$. It is then straightforward to solve the planner's Hamiltonian to arrive at the Euler equation

$$\frac{\dot{c}(t)}{c(t)} = \frac{A-\delta-\rho}{\sigma}, \tag{13.19}$$

the resource constraint $\dot{K}(t) = AK(t) - e^{nt}c(t) - \delta K(t)$, and the transversality condition

$$\lim_{T\to\infty} \left\{ c(T)^{-\sigma} e^{-\rho T} K(T) \right\} = \lim_{T\to\infty} \left\{ c(0)^{-\sigma} e^{-(A-\delta)T} K(T) \right\} = 0. \tag{13.20}$$

These expressions characterize the planner's solution in the AK model.

Interestingly, consumption grows at a constant rate irrespective of the level of the initial capital stock (see (13.19)). Hence, the AK model features no transition dynamics and the optimal consumption rule is $c(t) = (A-\delta-n-(A-\delta-\rho)/\sigma)e^{-nt}K(t)$. Capital likewise grows right away at the constant rate $\dot{K}(t)/K(t) = n + (A-\delta-\rho)/\sigma$.[24] The planner's solution could also be decentralized with competitive markets.

The AK model is an endogenous growth model as the rate of long-run growth is no longer an exogenous constant but a function of preferences and technology parameters and does respond to policies.[25] The reason why endogenous sustained growth is feasible in this economy is because there is no diminishing marginal product of capital. The chapter on the Solow model showed that a constant saving rate then leads to an (endogenous) balanced growth path (see Figure 3.2).

The production function that is linear in capital violates the standard diminishing returns assumption, and implies no role for labor to play in production. One would need massive increasing returns to scale (internally or externally), as in Romer (1986), to reconcile linear production in capital with the empirical capital share on the order of 1/3. Related, labor is very much an important factor in production that commands about 2/3 of all income. So the AK model is knife-edge and unrealistic in several dimensions. Still, it is helpful to discuss it as some more sophisticated theories of endogenous growth (to be presented shortly) have a reduced form AK structure.

[24] As always, it is easy to show that this solution fulfills the system of equations plus the transversality condition (given we imposed $\sigma(A-\delta-n) > A-\delta-\rho$). One can also show that this is the only solution. As consumption grows at a constant rate we can plug this into the constraint to get $\dot{K}(t)/K(t) = A - \delta - e^{(n+(A-\delta-\rho)/\sigma)t}c(0)$. This differential equation has a closed form solution given by $K(t) = \Xi e^{(A-\delta)t} + (A-\delta-n-(A-\delta-\rho)/\sigma)^{-1} e^{(n+(A-\delta-\rho)/\sigma)t}c(0)$, where Ξ is a constant of integration. The transversality condition is then only fulfilled if the constant of integration Ξ is equal to zero (which coincides with the solution above where capital grows at a constant rate).

[25] For instance, consider a decentralized version of this economy, where capital income is taxed at rate τ and the tax revenues are rebated lump-sum to the household. It is straightforward to show that the resulting growth rate is monotonically decreasing in τ.

Perhaps more realistically, the AK model can be viewed as a reduced form of a model with constant returns to physical and human capital combined.[26] That is, $Y(t) = F(K(t), H(t))$ where $H = hL$ is total human capital which can be accumulated at a steady rate without a bound. Such an economy may support a balanced growth path along which output, physical capital, and human capital all grow at a constant endogenous rate. But can human capital grow at a steady rate? The answer appears to be yes if there is increasing investment in human capital, such as from rising years of schooling. Much less clear, however, is whether human capital grows if the level of human capital investment per person is fixed over time. Each generation arguably starts at the same level of human capital, so they would need to obtain more human capital with each year of education or experience. Mincerian returns to education and experience exhibit no such secular trends.

Bils and Klenow (2000) argued that higher schooling has a level effect rather than a growth effect in a panel of counties. Klenow and Rodriguez-Clare (2005) also made a case for level effects of schooling. For example, they pointed out that education differences are quite persistent across countries, while income growth rate differences are more transitory. And schooling levels are generally trending up over time, while growth rates are not.[27] In contrast, schooling levels *are* highly correlated with income levels across countries.[28]

A property of all such AK-type endogenous growth theories is that they are entirely factor accumulation based. This means that, if inputs are measured properly, a growth accounting exercise should reveal zero TFP growth. These class of models also predicts a rather elastic response of the long-run growth rate to changes in taxes on profits or R&D subsidies that are not seen in the data (see Stokey and Rebelo, 1995). There is no role for patents, R&D expenditures, ideas, or purposeful innovation. We next study models which do provide a such role.

13.5.1 Endogenous growth through expanding varieties

Romer (1990) put the non-rival nature of ideas at center stage and modeled purposeful R&D investment of profit maximizing firms in general equilibrium. We discuss here a version of Romer (1990) that abstracts from physical capital.

Suppose the aggregate production function is given by

$$Y(t) = \frac{A}{1-\phi} L_y(t)^{\phi} \int_0^{N(t)} x(\nu, t)^{1-\phi} \, d\nu. \tag{13.21}$$

This aggregate production function is defined over the inputs labor L_y and an interval of different "machines" of measure N with quantity $x(\nu)$ for machine ν. We use the terminology

[26] See Appendix 13.A.4 where we comment on other purely accumulation based endogenous growth theories that look in a reduced form like the simple AK model.

[27] This is akin to Jones (1995) noting that R&D investments trend up over time but the long-run growth rate does not. He inferred that this likewise suggested a level effect of R&D investment on productivity.

[28] Imagine the law of motion of human capital $\dot{h} = Bh^{\phi}s^{\omega}$. The key question is whether $\phi = 1$ or $\phi < 1$. If $\phi = 1$ then higher levels of human capital investment s lead to perpetually faster growth. If, instead, $\phi < 1$ then higher levels of s yield higher levels of h but no faster growth in the long run. Again, this is analogous to semi-endogenous growth due to R&D investment that we will discuss further below.

machine even though, for simplicity, the inputs are like material in that they depreciate fully after use. We assume $\phi \in (0,1)$.

The production function (13.21) merits some discussion: First, the factor $1/(1-\phi)$ is added for notational simplicity and is not essential (one can always just redefine the constant A). Second, the production function has constant returns to scale when considering labor L_y and all the machines $\{x(\nu)\}_{\nu=0}^{N}$—in line with a simple replication argument. However, here the number of machine varieties N will be endogenous and will grow over time. Including this endogenous N, equation (13.21) features overall *increasing returns* to scale.[29] Third, one can view the production function as Cobb-Douglas over labor and a machine composite X, i.e., $Y = \frac{A}{1-\phi} L_y^{\phi} X^{1-\phi}$. The bundle of machines is a CES (Dixit-Stiglitz) aggregator

$$X = \left(\int_0^N x(\nu)^{\frac{\epsilon-1}{\epsilon}} d\nu \right)^{\frac{\epsilon}{\epsilon-1}}.$$

One arrives at (13.21) if one directly connects the elasticity of substitution across machines, ϵ, to the output elasticity of labor, ϕ, by setting $\epsilon = 1/\phi$. One can relax this assumption and still obtain a balanced growth path in the environment with research labor we study below.

Below we will refer to N as the knowledge stock in the economy. In this model knowledge is embodied in the number of available intermediate inputs. An economy with a higher knowledge stock will exhibit higher TFP.

Market structure The final output good is competitively produced by a representative firm. At each point in time this firm solves (so we suppress the time index)

$$\max_{L_y, \{x(\nu)\}_{\nu=0}^{N}} \left\{ \frac{A}{1-\phi} L_y^{\phi} \int_0^N x(\nu)^{1-\phi} d\nu - w L_y - \int_0^N p(\nu) x(\nu) d\nu \right\}, \tag{13.22}$$

where we chose the price of the final output as the numéraire. The first-order conditions of this problem are

$$\phi Y / L_y = w \tag{13.23}$$

and

$$A L_y^{\phi} x(\nu)^{-\phi} = p(\nu), \ \forall \nu. \tag{13.24}$$

In contrast to the final good, each machine variety is produced by only one firm, which therefore acts as a monopolistic competitor. The producer of a given variety ν maximizes current period profits

$$\pi(\nu) = \max_{p(\nu), x(\nu)} \{ p(\nu) x(\nu) - \psi x(\nu) \}, \text{ subject to (13.24)}. \tag{13.25}$$

Here the assumption is that machines can be produced at constant marginal cost ψ in units of the numéraire (final output good).

[29] To see this, suppose all machines cost p_x and a budget of E is equally split on all machines. We then have $x(\nu) = E/(p_x N)$, $\forall \nu$ and aggregate output becomes $Y = \frac{A}{1-\phi} L_y^{\phi} N^{\phi} (E/p_x)^{1-\phi}$. This highlights the love-of-variety inherent in specification (13.21): output is boosted if more types of machines are available (holding prices and the total budget constant).

Inventing varieties By hiring $1/(\eta N(t))$ labor units for R&D an entrant can invent a new machine. This implies for the law of motion of available varieties

$$\dot{N}(t) = \eta N(t) L_r(t), \text{ given } N(0) > 0, \tag{13.26}$$

where L_r is total labor used in R&D and $\eta > 0$. The N term in (13.26) represents a knowledge spillover: more available varieties make researchers more productive in terms of generating new varieties.[30]

After invention the entrant receives a perpetual patent to exclusively produce the invented variety. The value of such a patent is given at date t by the present discounted value of future profits

$$V(t) = \int_t^\infty e^{-\int_t^s r(\tau)d\tau} \pi(s) ds, \tag{13.27}$$

where r is the real interest rate and we suppressed the variety index ν due to symmetry. By time differentiating (13.27) one obtains the Hamilton-Jacobi-Bellman (HJB) equation: $V(t)r(t) = \pi(t) + \dot{V}(t)$. The return on V must be equal to the flow of profits $\pi(t)$ plus capital gains $\dot{V}(t)$. The stream of future profits in (13.27) incentivize the R&D investments in the first place. Under free entry in R&D we can assume that there is a representative R&D firm solving in each point in time

$$\max_{L_r} \{V(t)\eta N(t) L_r(t) - w(t) L_r(t)\} \tag{13.28}$$

subject to

$$L_r(t) \geq 0.$$

Household problem The model is closed with a standard household side. The representative household supplies inelastically L units of labor earning a wage rate w. The household owns all the firms, whose total value represents household wealth

$$a(t) = \int_0^{N(t)} V(\nu, t)\, d\nu, \tag{13.29}$$

that earns a combined real return of r. The household then solves

$$\max_{\{a(t), c(t)\}_{t=0}^\infty} \int_0^\infty e^{-\rho t} \frac{c(t)^{1-\sigma} - 1}{1-\sigma} \tag{13.30}$$

subject to

$$\dot{a}(t) = r(t)a(t) + w(t)L - c(t),$$

and a standard no-Ponzi game condition.[31] We deliberately abstract from population growth for now; more on that below.

A decentralized equilibrium is defined as a path of prices and quantities that jointly solve the household and the firm problem, and is consistent with factor market clearing. Appendix

[30] We will relax the one-for-one proportionality of this spillover in N below.

[31] The no-Ponzi game condition can be expressed as $\lim_{T\to\infty} \left\{ e^{-\int_0^T r(\tau)d\tau} a(T) \right\} \geq 0$.

13.A.5 gives a formal equilibrium definition. To ensure that the household problem is well defined (i.e., utility is bounded) we need to impose the parameter restriction

$$\rho > (1-\sigma)\frac{(1-\phi)\eta L - \rho}{1-\phi+\sigma}. \tag{13.31}$$

This condition ensures that the discount rate is larger than the endogenous consumption growth rate times $1-\sigma$. Furthermore, we focus in the following on the interesting case with strictly positive growth and $L_r > 0$, which will be ensured as long as

$$\rho < (1-\phi)\eta L. \tag{13.32}$$

This condition is fulfilled and innovations are profitable as long as the R&D efficiency η and the market size L are large enough compared to the discount rate.

Note that the decentralized equilibrium we characterize involves monopolistically competitive firms and knowledge spillovers. As a consequence, as we will see below, the decentralized equilibrium will not coincide with the planner's solution.

Solving the model We already solved the final output producer problem above, yielding the first-order conditions (13.23) and (13.24). We did so in order to specify the firm problem of the machine producers who take the (inverse) demand (13.24) into account.

Solving the problem of a machine producer gives the first-order conditions for any point in time

$$x(\nu) = \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1}{\phi}} L_y, \ \forall \nu, \tag{13.33}$$

and the resulting optimal price is given by $p(\nu) = \psi/(1-\phi)$, i.e., a constant markup over marginal cost.[32] Profits are given by

$$\pi(\nu) = \phi A \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1}{\phi}-1} L_y. \tag{13.34}$$

This model features a market size effect, i.e., profits from having developed a new machine variety are increasing in the number of workers, L_y, operating this machine. This force plays a role for the direction of technical change in the presence of high and low skilled labor in Acemoglu (1998). See Chapter 21 for further discussion.

Because of symmetry—$x(\nu)$ and $p(\nu)$ are the same for all ν—we can suppress the variety index ν and write final output (13.21) as

$$Y(t) = \frac{A}{1-\phi}\left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} L_y(t) N(t), \tag{13.35}$$

and the wage rate as

$$w(t) = \phi\frac{A}{1-\phi}\left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} N(t). \tag{13.36}$$

[32] The optimal price fulfills the Lerner condition that price is equal to $\frac{\partial x}{\partial p}\frac{p}{x}/\left(\frac{\partial x}{\partial p}\frac{p}{x}+1\right)$ times marginal cost. Under the imposed CES structure the price elasticity of demand $\frac{\partial x}{\partial p}\frac{p}{x}$ is equal to the constant $-1/\phi$. The markup is then $(1/\phi)/(1/\phi-1)$. This markup rule is familiar from Chapter 6.

Plugging (13.33) and (13.35) into the final good market clearing condition gives consumption as

$$c(t) = \left(\frac{1}{(1-\phi)^2} - 1\right) \psi \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1}{\phi}} L_y(t) N(t). \tag{13.37}$$

The household problem gives rise to the same Euler equation as before:

$$\frac{\dot{c}(t)}{c(t)} = \frac{r(t) - \rho}{\sigma}, \tag{13.38}$$

as well as the transversality condition

$$\lim_{T\to\infty} \left\{ e^{-\int_0^T r(\tau)d\tau} a(T) \right\} = 0.$$

The first-order condition for the R&D firm's problem is given by $V(t) \leq \frac{w(t)}{\eta N(t)}$, with $L_r(t)(V(t) - w(t)/(\eta N(t))) = 0$ (taking care of the Kuhn-Tucker condition). In the case with strictly positive growth and R&D investment—which we focus on in the following—a firm value then needs to equalize the fixed cost of inventing a variety:

$$V(t) = \frac{w(t)}{\eta N(t)}. \tag{13.39}$$

Substituting in the wage from (13.36), this implies

$$V(t) = \frac{\phi A}{\eta(1-\phi)} \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}}, \ \forall t, \tag{13.40}$$

i.e., a *constant* equilibrium firm value over time. Substituting this constant firm value and the equilibrium profits into the HJB equation then gives

$$r(t) = \frac{\pi(t)}{V(t)} = \eta(1-\phi) L_y(t). \tag{13.41}$$

Finally, we obtain from substituting the labor market clearing condition in the law of motion of N

$$\frac{\dot{N}(t)}{N(t)} = \eta(L - L_y(t)). \tag{13.42}$$

This system of equations can be simplified. Using (13.37), (13.41), and (13.42) in the Euler equation, we obtain a single first-order differential equation in L_y given by

$$\frac{\dot{L}_y(t)}{L_y(t)} = \frac{\eta(1-\phi+\sigma)L_y(t) - \rho}{\sigma} - \eta L. \tag{13.43}$$

Hence, there is indeed a balanced growth path along which L_y is constant with

$$L_y^\star = \frac{\sigma L + \rho/\eta}{1-\phi+\sigma}. \tag{13.44}$$

There are no transitional dynamics and the economy will grow right away along the balanced growth path (similar to the AK model above).[33] R&D labor along the balanced growth path is given by[34]

$$L_r^\star = \frac{(1-\phi)L - \rho/\eta}{1-\phi+\sigma}.$$

The interest rate along the balanced growth path is then

$$r^\star = (1-\phi)\frac{\eta\sigma L + \rho}{1-\phi+\sigma}. \tag{13.45}$$

The number of varieties, output, consumption, and wealth grow at rate

$$\frac{\dot{N}(t)}{N(t)} = \frac{\dot{Y}(t)}{Y(t)} = \frac{\dot{c}(t)}{c(t)} = \frac{\dot{a}(t)}{a(t)} = \frac{(1-\phi)\eta L - \rho}{1-\phi+\sigma}. \tag{13.46}$$

The transversality condition is ensured by (13.31).

We see that the decentralized growth rate is decreasing in the preference parameters ρ and σ.[35] The endogenous growth rate (13.46) is increasing in the labor force L, a so-called strong scale effect. Given an R&D intensity, $L_r^\star/L$, profits are increasing in scale, thereby attracting more R&D investment. Furthermore, the share of labor allocated to R&D $L_r^\star/L$ is increasing in L. When applied to closed economies, the model predicts bigger countries will grow faster and exhibit higher R&D intensity. These predictions are clearly not borne out in the data (neither in the cross-section of countries not in the time series within a country). The section below which limits the knowledge spillovers will address this critique.

Due to the increasing overall returns to scale in the production function (13.21) and the monopolistically competitive firms the decentralized equilibrium is not efficient. The planner solves:

$$\max_{\{L_y(t),c(t),\{x(\nu,t)\}_{\nu=0}^{N}(t),N(t)\}_{t=0}^{\infty}} \int_0^\infty e^{-\rho t}\frac{c(t)^{1-\sigma}-1}{1-\sigma}dt, \tag{13.47}$$

subject to the resource constraint

$$\frac{A}{1-\phi}L_y(t)^\phi \int_0^{N(t)} x(\nu,t)^{1-\phi}\,d\nu = c(t) + \int_0^{N(t)} \psi x(\nu,t)\,d\nu, \tag{13.48}$$

[33] Transition dynamics do not arise because the knowledge spillovers in (13.26) are proportional to N and because there is no population growth. We give up both assumptions below. One can directly see from (13.43) that this balanced growth path is the only equilibrium path. For any initial $L_y < L_y^\star$, L_y will shrink over time and go to zero and so does r and the transversality condition is violated. For any initial $L_y > L_y^\star$, L_y will systematically grow over time and go to L implying a "high" interest rate in (13.41), and yet no further growth because there is zero research labor, and in turn an inconsistency with the Euler equation.

[34] In the analysis above we considered an interior solution with $L_r > 0$. Can there also be a corner solution with $L_r = 0$? In such a case there is no growth and therefore $r^\star = \rho$, implying firm value

$$V = \pi/r^\star = \phi A\left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1}{\phi}-1} L,$$

which under (13.32) is strictly larger than $w(t)/\eta N(t)$. Hence, the condition $V(t) \leq w(t)/\eta N(t)$ is violated and (13.32) ensures an equilibrium with strictly positive growth.

[35] Intuitively, with more impatience or more curvature in the momentary utility function the consumption path is less steep.

and the ideas production function (13.42) and a given $N(0) > 0$. (Note that there we already used the labor constraint to substitute out $L_r(t)$.) For the solution to this problem to be well defined and leading to positive growth, we need to impose the parameter restrictions

$$(1-\sigma)\eta L < \rho < \eta L. \tag{13.49}$$

In Appendix 13.A.6 we solve the planner problem and show that it implies a balanced growth path with

$$g^{SP} = \frac{\eta L - \rho}{\sigma}, \tag{13.50}$$

and that the quantity of each machine is given by

$$x(\nu,t) = \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} L_y^{SP} = \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{(\sigma-1)L + \rho/\eta}{\sigma}. \tag{13.51}$$

Again, one can also show that there are no transitional dynamics and the economy grows right away along the balanced growth path (irrespective of the initial $N(0)$).

In the decentralized equilibrium above we saw that $x(\nu,t) = (A(1-\phi)/\psi)^{\frac{1}{\phi}} L_y^\star$. Hence, for given production labor L_y the decentralized machine quantity is too low by a factor of $(1-\phi)^\phi < 1$. This is due to the monopolistically competitive firms charging a markup factor of $1/(1-\phi)$ over marginal cost. This problem of under-usage of machines could be fixed by subsidizing machines by a rate τ that fulfills $1-\tau = (1-\phi)^{-1}$ and finance this subsidy by lump-sum taxes. However, even such a subsidy would not restore overall efficiency in this economy. This is because on top of the distortion due to markups there is the knowledge spillover in (13.42) that the planner takes into account, whereas in the decentralized equilibrium the R&D effort is only determined by the (discounted stream of) profits. Contrasting the decentralized equilibrium with the planner solution we indeed see that the decentralized growth rate is generally too low. As a consequence, restoring efficiency requires an additional R&D subsidy on top of the machine subsidy.

It is straightforward to add physical capital to the endogenous growth model above where additional transitional dynamics can arise due to the convergence of the capital stock to its steady state level.

The endogenous growth theory above is micro-founding the forward looking R&D decision of monopolistically competitive firms and characterizes how it determines the long-run growth rate of the economy. Markups serve a double role; they lead to efficiency losses due to too high machine prices, but they also lead to profits and incentivize R&D expenditures. The model however remains rather stylized in many dimensions and the resulting policy conclusions are rather stark. Along a balanced growth path the HJB equation, (13.39) and (13.36) imply

$$\frac{\pi^\star}{r^\star} = \frac{\phi}{\eta}\frac{A}{1-\phi}\left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}}. \tag{13.52}$$

Hence, the ratio of profits and the interest rate is equal to a constant. Taxing profits would lower the interest rate to therefore lower the growth rate (see (13.38)). Similarly, lowering the markup firms can charge to $\mu < (1-\phi)^{-1}$ through say antitrust policy would monotonically lower the growth rate. Or making patents of the innovators expiring would also systematically decrease the growth rate. The quality ladder model we discuss below—which features business stealing—can imply more nuanced and non-monotonic policy trade-offs.

Less than proportional knowledge spillovers in ideas production

Here we introduce two modifications to the framework above introduced by Jones (1995). First, suppose instead of (13.26) that the "ideas" production function is given by

$$\dot{N}(t) = \eta N(t)^{\epsilon} L_r(t), \text{ given } N(0) > 0, \tag{13.53}$$

with $\epsilon < 1$ capturing that it requires—as the knowledge stock advances—more and more R&D labor to sustain a constant rate of knowledge growth. With $0 < \epsilon < 1$ there are still positive (but limited) knowledge spillovers, i.e., with a higher knowledge stock, N, each R&D worker generates more ideas. In contrast, with $\epsilon < 0$ the specification can also capture a "fishing out" effect (negative technology spillovers) where it gets harder for a given R&D worker to find a new idea the higher is the knowledge stock. The special case with $\epsilon = 0$ captures a situation without any knowledge spillovers and with $\epsilon = 1$ we are back in the framework above with proportional knowledge spillovers. Second, suppose further that there is population growth at rate $n > 0$ such that the labor market clearing condition becomes $L(t) = Le^{nt} = L_y(t) + L_r(t)$. To ensure that the household problem is bounded we assume $\rho - n > n/(1-\epsilon)$.

With these two modifications to the set-up, how do the equilibrium dynamics look and is there still a balanced growth path along which all variables grow at constant rates? It is important to note that—as there is population growth—L_y and L_r both have to grow at rate n along a balanced growth path (if not at least one of L_y or L_r cannot change at a constant rate). The flow profits are still given by (13.34) which are now however time dependent as L_y is growing. Given a constant interest rate the value of a patent at time t along the balance growth path is then given by (see (13.27))

$$V(t) = \frac{\pi(t)}{r^{\star} - n} = \phi A \left(\frac{A(1-\phi)}{\psi} \right)^{\frac{1}{\phi}-1} \frac{L_y(t)}{r^{\star} - n}. \tag{13.54}$$

Again this value is now time dependent as the market size is expanding due to population growth. In an equilibrium with positive growth $V(t)$ has to equalize the marginal cost of innovating which is given by $\frac{w(t)}{\eta N(t)^{\epsilon}}$ under the new spillover specification. Substituting in the wage rate (which is still given by (13.36)) the optimality condition of the representative R&D firm reduces to

$$\frac{L_y(t)}{r^{\star} - n} = \frac{1}{\eta(1-\phi)} N(t)^{1-\epsilon}. \tag{13.55}$$

Can this equation be fulfilled for all t? By differentiating with respect to time we see the answer is yes if and only if

$$\frac{\dot{N}(t)}{N(t)} = \frac{n}{1-\epsilon}. \tag{13.56}$$

Then per-capita consumption, wages and per-capita output all grow at this rate $\frac{n}{1-\epsilon}$ and the interest rate $r^{\star} = \rho + \sigma n/(1-\epsilon)$ is pinned down by the Euler equation. It is straightforward to check that the transversality condition is indeed fulfilled along this path.

This model version is sometimes called a semi-endogenous growth model because the long-run growth rate is just determined by two parameters: the population growth rate n

and a technology parameter ϵ. Intuitively, larger knowledge spillovers (a higher ϵ) increases the growth rate. With $\epsilon < 1$ the ideas production function runs into decreasing returns and the only way to sustain long-run growth is by pushing with an increasing number of researchers over time. Hence, without population growth, i.e., with $n = 0$, sustained long-run growth is not feasible. The long-run growth rate is efficient and there is no role for policy to influence it.[36] Hence, the model does not have rich predictions about how policy and the market structure interact with the long-run growth rate. But the model's prediction is in line with the time series data from the U.S. or the OECD countries that shows a rather stable productivity growth while population (and the number of researchers) is growing at a steady rate. Another strength of the model is that it adds demographics to the potential determinants of long-run growth (see Peters and Walsh (2023) for an application that links population growth to firm entry rates and growth).

In contrast to the model with proportional knowledge spillovers we studied before, it is important to note that this model now features transitional dynamics. Substituting the growth rate (13.56) into (13.53) and using the labor market clearing gives

$$\frac{n}{\eta(1-\epsilon)} = N(t)^{\epsilon-1}(Le^{nt} - L_y(t)), \tag{13.59}$$

which together with (13.55) specifies for given population Le^{nt} a unique $N(t)$ that supports the balanced growth. Whenever the initial N deviates from this level there will be transitional dynamics. Furthermore, as ϵ may be relatively close to one the transitional dynamics may be slow (unlike the convergence of the physical capital stock in the neoclassical growth model).[37] Policy will affect the long-run output level in this framework. Equation (13.59) implies along the balanced growth path $N(t) = (\eta(1-\epsilon)L_r(t)/n)^{\frac{1}{1-\epsilon}}$. Substituting this $N(t)$ into (13.35) allows us to express total output per worker in final goods production along the balanced growth path as

$$\frac{Y(t)}{L_y(t)} = \mathcal{A}\left(\frac{L_r(t)}{L(t)}\right)^{\frac{1}{1-\epsilon}} L(t)^{\frac{1}{1-\epsilon}}, \tag{13.60}$$

where $\mathcal{A}$ is some constant. Hence, in this model there is no scale effect on the long-run growth rate but for a given research intensity $L_r(t)/L(t)$ the model predicts higher per-capita income *levels* in larger economies. Similarly, as policy can affect research intensity it can affect the long-run *level* of income.

In the endogenous growth framework studied in this section, all long-run growth is generated by an expanding (input) variety. Though this is stark for exposition, there is indeed

[36] The planner problem is given by

$$\max_{\{L_y(t),c(t),\{x(\nu,t)\}_{\nu=0}^{N}(t),N(t)\}_{t=0}^{\infty}} \int_0^\infty e^{-(\rho-n)t}\frac{c(t)^{1-\sigma}-1}{1-\sigma}dt, \tag{13.57}$$

subject to

$$\frac{A}{1-\phi}L_y(t)^\phi \int_0^{N(t)} x(\nu,t)^{1-\phi}d\nu = c(t) + \int_0^{N(t)} \psi x(\nu,t)d\nu, \tag{13.58}$$

$\dot{N}(t) = \eta N(t)^\epsilon(Le^{nt} - L_y(t))$ and a given $N(0) > 0$. In the Appendix 13.A.7 we solve this planner problem and show that its solution also supports a balanced growth path with $\frac{\dot{N}(t)}{N(t)} = \frac{n}{1-\epsilon}$.

[37] Bloom, Jones, and Van Reenen (2020) estimate for example $\epsilon = 0.8$ in the semiconductor industry.

evidence that variety is growing and contributing to growth. Broda and Weinstein (2006) find that the U.S. import growth comes from a rising number of country-product pairs over time, across final consumer goods, intermediate goods, and business equipment. Such evidence breathes empirical life into influential models of trade in varieties such as Krugman (1980), Rivera-Batiz and Romer (1991), and Melitz (2003).

Broda and Weinstein (2010) document an expanding set of products at U.S. grocery stores. Related, Hottman, Redding, and Weinstein (2016) show that about one-third of variation in firm size and growth can be traced to the number of distinct consumer packaged goods in their portfolio. Hsieh and Rossi-Hansberg (2023) document that major retailers and service producers have opened more establishments in the U.S., especially since 2000. Given that such products and services are often nontradable, new locations offer local consumers growing choice. Hsieh and Rossi-Hansberg emphasize that such rival investments in local establishments are distinct from the nonrival corporate investments in products, services, and store layout. Garcia-Macia, Hsieh, and Klenow (2019), Klenow and Li (2021), and Peters and Walsh (2023) all estimate that between one-fifth and one-third of U.S. growth comes from rising variety.

13.5.2 Growth through quality ladders and creative destruction

The Romer model sketched out in the previous subsection endogenizes innovation, but entirely through new varieties. It does not feature any quality improvements in products after they are invented. Sales and profits of existing varieties are unaffected by the arrival of "horizontal" new varieties.

In contrast, it seems routine to observe companies improving their products over time. Think of successive generations of car models and microprocessors. The Romer model also abstracts from product and firm exit due to "creative destruction." Schumpeter's vision, formalized by Aghion and Howitt (1992), features innovation by new producers that displaces closely substitutable products of existing producers. That is, "vertical" innovation that destroys the sales and profits of existing producers. Figure 13.6 shows annual firm exit rates from the U.S. Census Business Dynamics Statistics (BDS) database for 1980–2022. Over 10% of small firms (1–4 employees) exit in a typical year, and around 2% of firms with 20 or more employees.

There is also substantial job reallocation among surviving establishments, a fact first documented by Davis, Haltiwanger, and Schuh (1998). The job reallocation rate is the sum of the job creation rate and the job destruction rate minus the absolute difference between the two. The job creation rate is the sum of employment gains at surviving and opening establishments from year $t-1$ to year t divided by average aggregate employment across the two years. The job destruction rate is defined similarly, only with a numerator equal to the sum of employment losses at surviving and exiting establishments from $t-1$ to t. Figure 13.7 plots the annual rate of job reallocation for 5-year periods from 1980–2019 in the BDS. It averages 20–30%, but is on a downward trajectory. Job reallocation occurs mostly within narrow industries, so it reflects competition among close competitors rather than broad sectoral shifts such as from goods to services. Garcia-Macia et al. (2019) argue that fitting such high exit and job destruction rates requires a prominent role for creative

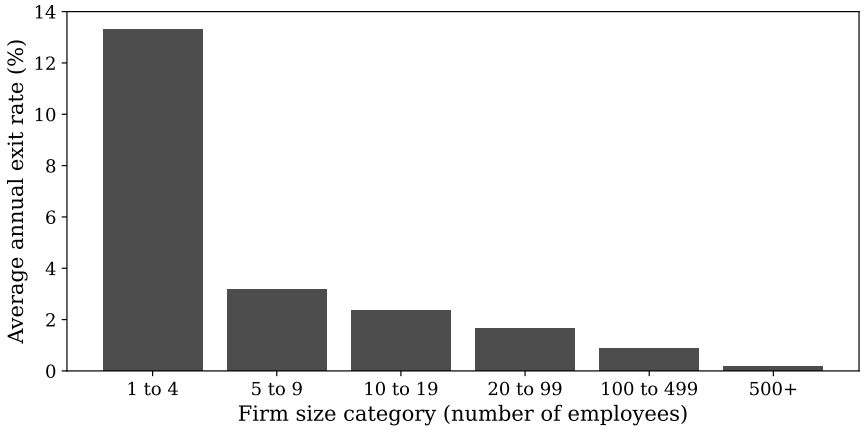


Figure 13.6: Firm exit rates in the U.S., 1980–2022

Source: U.S. Business Dynamics Statistics (BDS).

destruction in overall growth.[38]

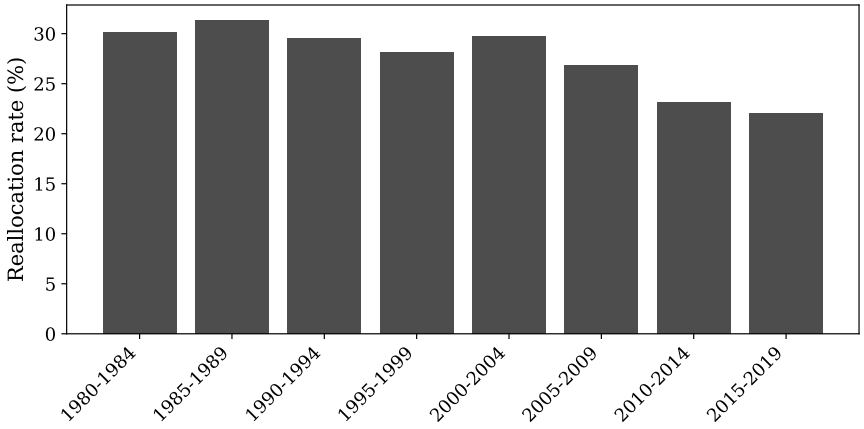


Figure 13.7: Job reallocation in the U.S., 1980–2019

Source: U.S. Business Dynamics Statistics (BDS).

Figure 13.8 compares the average size of new firms (age 0), firms age 1–5 years, 6–10 years, and 11+ years in the U.S. Young firms have 5–10 workers on average, whereas the oldest firms average 35–40 workers. Hsieh and Klenow (2014) report that this pattern stems equally from selection (smaller firms are more likely to exit) and survivor growth. Figure 13.9 plots the employment share of entrants (defined as firms aged 0 years) in the U.S. over 5-year periods from 1980–2019. Figure 13.10 plots the employment share of exiting firms in the U.S. over 5-year periods from 1980–2019. Their share fell from a bit over 3% to about 2%. The combination of falling job reallocation, entry, and exit rates has generated concern that declining "business dynamism" has contributed to slower productivity growth in the U.S. in recent decades. Decker, Haltiwanger, Jarmin, and Miranda (2016) and Akcigit and

[38] Chapter 22 presents the time-series patterns of establishment entry and exit, as well as the job creation and job destruction rates.

Ates (2023) present related evidence and a model.

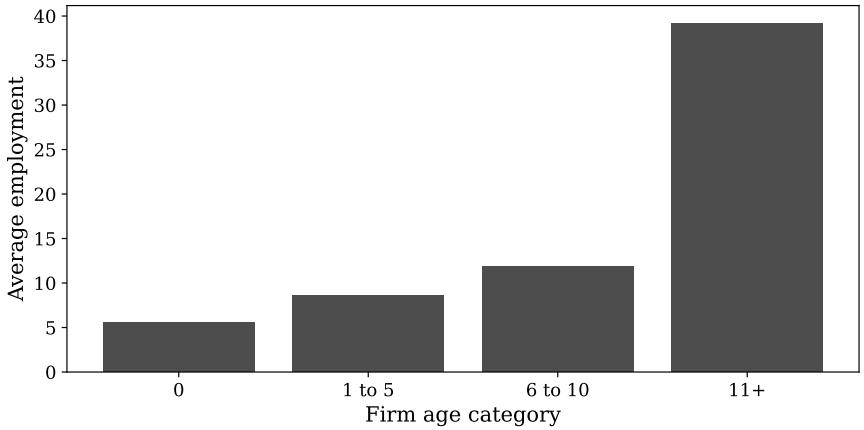


Figure 13.8: Firm size and firm age, 1988–2022

Source: U.S. Business Dynamics Statistics (BDS).

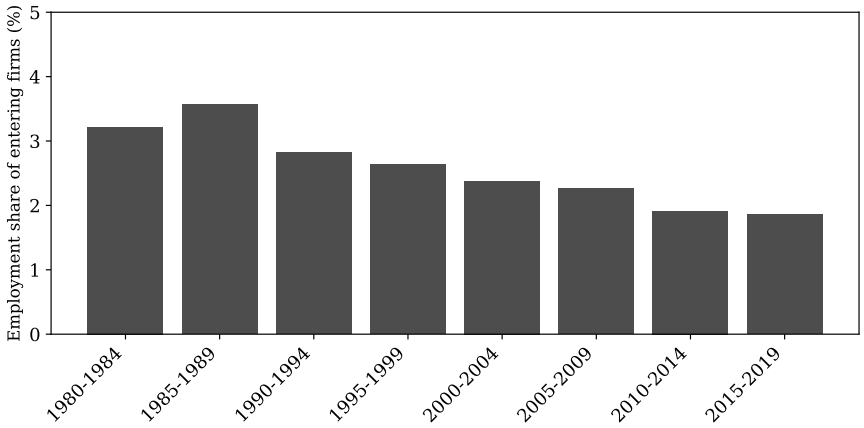


Figure 13.9: Entrant employment share in the U.S., 1980–2019

Source: U.S. Business Dynamics Statistics (BDS).

Finally, Figure 13.11 plots industry productivity growth against industry exit rates, both averaged over 5-year periods within 1988–2022. This involves 160 sectors and 7 5-year periods for a total of 1,104 sector-periods. The plot displays 20 bin-scatter points. Consistent with theories of creative destruction, industries with high exit rates tend to exhibit faster productivity growth. See Adhami (2025) for documentation of these facts and the use of entry and exit rates to infer knowledge spillovers from incumbents to entrants.

To speak to the substantial churn of market shares, even within narrow industries, we now sketch a model of growth through creative destruction.

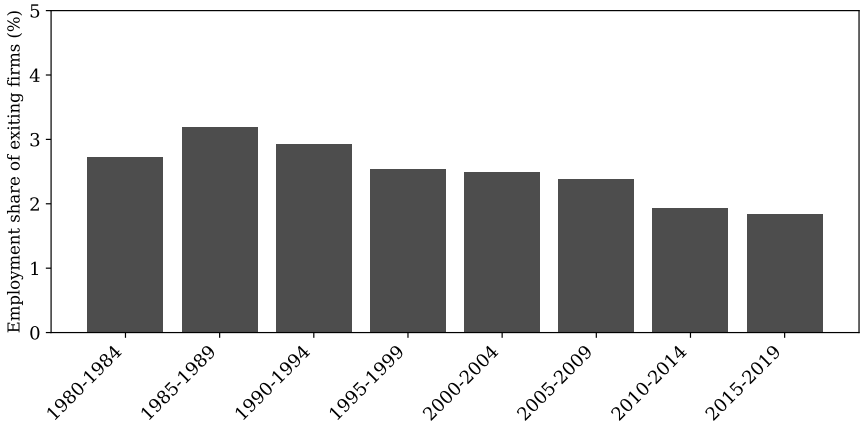


Figure 13.10: Exiting firm employment share in the U.S., 1980–2019

Source: U.S. Business Dynamics Statistics (BDS).

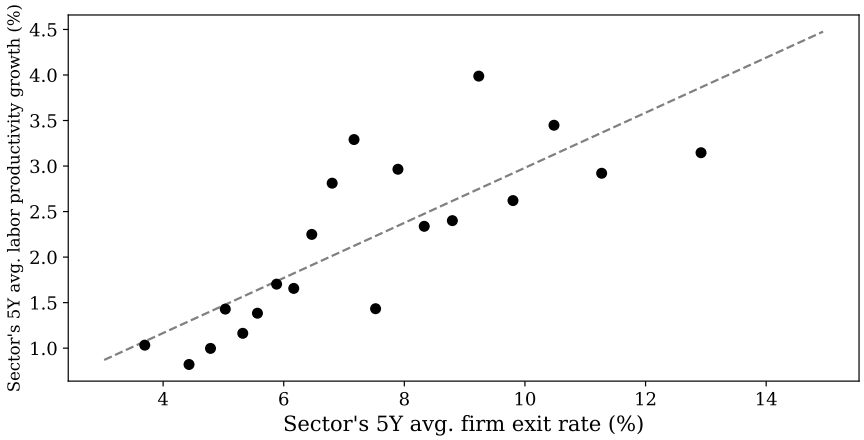


Figure 13.11: Growth and exit rates in the U.S., 1988–2022

Source: Labor productivity growth rates from the U.S. Bureau of Labor Statistics and exit rates from the U.S. Business Dynamics Statistics (BDS).

Quality Ladders model Final good producers use a fixed measure one of intermediate good varieties with evolving quality levels $q(\nu, t)$:

$$Y(t) = \frac{A}{1-\phi} L(t)^{\phi} \int_0^1 q(\nu, t) x(\nu, t)^{1-\phi} \, d\nu \tag{13.61}$$

where $0 < \phi < 1$ so that there is diminishing returns to any one intermediate line ν. A fixed amount of labor $L(t) = L$ can be used to produce final goods.

Output can be used for consumption, intermediate goods, and research:

$$Y(t) = C(t) + X(t) + Z(t).$$

This is sometimes called a "lab equipment" model because intermediate goods (which for simplicity depreciate fully here rather than being durable equipment) are used to make

research goods. Unlike in the Romer model, here there is no knowledge spillover from past innovation to current research. Intermediate goods, in turn, are produced according to

$$X(t) = \int_0^1 \psi q(\nu, t) x(\nu, t) d\nu,$$

so that one unit of intermediate good ν requires $\psi \cdot q(\nu, t)$ units of the final good.

Aggregate research effort is merely the sum of effort directed at improving each intermediate good's quality

$$Z(t) = \int_0^1 Z(\nu, t)\, d\nu.$$

Quality improves on a product line in discrete increments in proportion $\lambda > 1$:

$$q(\nu, t) = \lambda^{m(\nu,t)} q(\nu, 0) \quad \forall \nu, t.$$

See Figure 13.12. Here $m(\nu, t)$ is the cumulative number of quality steps taken on line ν between time 0 and t. The flow rate of innovations on product line ν at moment t is proportional to research effort on that line:

$$z(\nu, t) \equiv \eta \frac{Z(\nu, t)}{q(\nu, t)},$$

where $\eta > 0$. Note that it is more difficult to innovate on a line the higher the quality attained on that line.

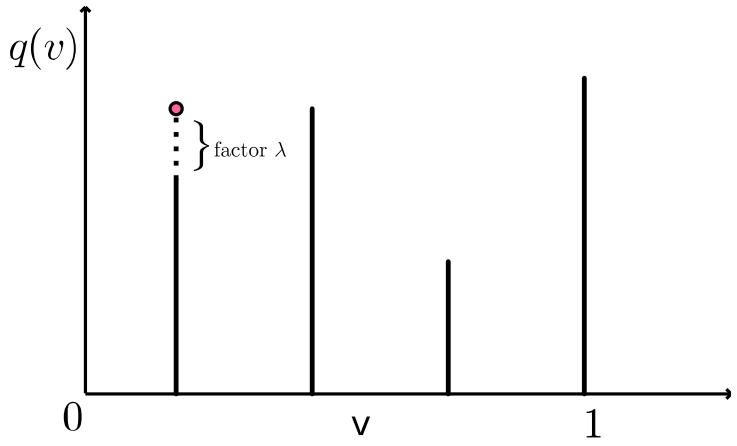


Figure 13.12: Quality ladders

Finally, the representative agent's utility function takes the standard CRRA form:

$$U(0) = \int_0^\infty e^{-\rho t} \frac{C(t)^{1-\sigma} - 1}{1 - \sigma} dt$$

Decentralized equilibrium The final goods sector is competitive and producers maximize current profits taking all prices as given (the final good whose price is normalized to

one, the intermediate good prices, and the wage):

$$\Pi(t) = \frac{A}{1-\phi} L(t)^{\phi} \int_0^1 q(\nu,t) x(\nu,t)^{1-\phi} d\nu - \int_0^1 p(\nu,t) x(\nu,t)\, d\nu - w(t)L(t).$$

The first-order conditions for this problem imply

$$w(t) = \phi \frac{Y(t)}{L}, \tag{13.62}$$

and

$$x(\nu,t) = \left[\frac{A\, q(\nu,t)}{p(\nu,t)}\right]^{1/\phi} L, \tag{13.63}$$

where we already used labor market clearing $L(t) = L$. The wage is equal to the marginal product of labor, and intermediate input demand is inversely related to a variety's *quality-adjusted* price. Intermediate demand is increasing in production labor, a complementary input.

Intermediate good producers are monopolists over a particular quality level. We assume innovations are "drastic" in that

$$\lambda \geq \left(\frac{1}{1-\phi}\right)^{\frac{1-\phi}{\phi}}, \tag{13.64}$$

so that producers are not constrained by competition from firms who can produce at lower rungs on the quality ladder for their variety. Instead, their closest competitors are other varieties.[39] With a continuum of varieties they are monopolistic competitors. They maximize current profits:

$$\pi(\nu,t) = p(\nu,t)x(\nu,t) - \psi q(\nu,t)x(\nu,t)$$

Intermediate good monopolists take downstream demand as given by (13.63) and choose prices and quantities. This results in the profit-maximizing price

$$p(\nu,t) = \frac{\psi}{1-\phi} q(\nu,t).$$

The marginal cost is $\psi \cdot q(\nu,t)$ and the markup is $1/\phi/(1/\phi - 1)$, as $1/\phi$ is the elasticity of substitution across intermediate good varieties in final goods production. Using the profit-maximizing price, input demand becomes

$$x(\nu,t) = \left(\frac{A(1-\phi)}{\psi}\right)^{1/\phi} L, \quad \forall\, \nu, t.$$

and in turn maximized profits are

$$\pi(\nu,t) = \phi A \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} q(\nu,t) L, \quad \forall\, \nu, t. \tag{13.65}$$

[39] See the Appendix 13.A.9 which derives condition (13.64) and illustrates how limit pricing might prevail in the case of "incremental" innovations with step sizes below the monopoly markup if condition (13.64) does not hold.

We can use the intermediate good quantities x (which do not depend on v or t) to find an expression for aggregate output:

$$Y(t) = \frac{A}{1-\phi}\left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} Q(t)L,$$

where $Q(t)$ is defined as the average quality of intermediate goods

$$Q(t) \equiv \int_0^1 q(\nu,t)\,d\nu.$$

In turn, the amount of aggregate output devoted to intermediate goods is

$$X(t) = \int_0^1 \psi q(\nu,t)x(\nu,t)d\nu = \psi^{1-1/\phi}(A(1-\phi))^{1/\phi}Q(t)L.$$

Meanwhile, there is free entry into the research sector. Research firms choose how much research to devote to each variety:

$$\max_{Z(\nu,t)} \frac{\eta Z(\nu,t)}{q(\nu,t)}\lambda V(\nu,t) - Z(\nu,t),$$

where $V(\nu,t)$ is the expected present discounted value of profits for the current leader. The innovator takes over the entire profit stream (creative destruction), and earns profits that are scaled up by $\lambda > 1$ relative to that of the previous leader because profits are proportional to the level of quality in (13.65). The first-order condition for researchers is

$$\frac{\eta}{q(\nu,t)}\lambda V(\nu,t) = 1.$$

Note that, if there is positive research on all varieties, then the ratio of the value of a variety to its quality level is the same for all varieties.

Would incumbents put effort into improving their own products to stave off creative destruction? If they did so, they would only reap the *increment* to their value, and this would not cover the cost of research to them:

$$\frac{\eta}{q(\nu,t)}(\lambda-1)V(\nu,t) < \frac{\eta}{q(\nu,t)}\lambda V(\nu,t) = 1.$$

This is called the **Arrow Replacement Effect**, which says that incumbents have less to gain from innovation than entrants because $\lambda - 1 < \lambda$.[40]

The expected present discounted value of profits from being the current leader of a quality ladder is:

$$V(\nu,t) = \int_t^\infty \left[e^{-\int_t^s r(s')ds'} \cdot e^{-\int_t^s z(\nu,s')ds'}\right]\pi(\nu,s)ds$$

[40]In order to explain why incumbents very much do improve their own products in the world, one could alter the model to have incumbents enjoy lower costs of innovating. If such costs are convex, however, then there could be both incumbent innovation on their own products and creative destruction of competitor products in equilibrium.

where $z(\nu, t)$ is the instantaneous rate of innovations on line ν, $z(\nu, t) \triangleq \eta \cdot Z(\nu, t)/q(\nu, t)$. The stream of profits is discounted by the real interest rate and the probability of creative destruction.

Finally, a representative household maximizes the presented discount value of its utility from consumption:

$$\max_{\{C(t), \mathcal{A}(t)\}_{t=0}^{\infty}} \int_0^{\infty} e^{-\rho t} \frac{C(t)^{1-\sigma} - 1}{1-\sigma} dt$$

subject to

$$\dot{\mathcal{A}}(t) = r(t)\mathcal{A}(t) + w(t)L - c(t), \, \forall t,$$

and a no-Ponzi game condition. Note that assets are simply the combined value of all intermediate good monopolists, who are owned by the household:

$$\mathcal{A}(t) \equiv \int_0^1 V(\nu, t) \, d\nu.$$

The current value Hamiltonian for the household is therefore

$$\mathcal{H}(c, a, \mu) = \frac{c^{1-\sigma} - 1}{1-\sigma} + \mu \left(r \cdot a + w - c \right)$$

where lower case c and a denote per-capita consumption and assets, respectively. The solution to this problem yields the standard Euler equation

$$\frac{\dot{c}}{c} = \frac{r - \rho}{\sigma},$$

and the usual transversality condition is

$$\lim_{T \to \infty} a(T) \exp \left[\int_0^T -r(s) ds \right] = 0.$$

As with the Romer model, one can show that the unique decentralized equilibrium of the quality ladder model entails no transition dynamics. That is, the economy is always on a BGP with growth rate

$$g^{\star} = \frac{\dot{Y}}{Y} = \frac{\dot{C}}{C} = \frac{\dot{X}}{X} = \frac{\dot{Q}}{Q}.$$

Note that, because L is constant, $c \equiv C/L$ grows at the same rate as C does. The Euler equation then gives us one equation in the two unknowns

$$r^{\star} = \sigma g^{\star} + \rho,$$

where $r^{\star}$ is the BGP real interest rate. And free entry into research gives us

$$\frac{V(\nu, t)}{q(\nu, t)} = \frac{1}{\lambda \, \eta}, \quad \forall \, \nu, t.$$

To obtain a constant value of each variety relative to its quality, we must have $z(\nu,t) = z^\star, \ \forall\, \nu, t$. This in turn implies

$$\frac{V(\nu,t)}{q(\nu,t)} = \frac{\phi A \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} L}{r^\star + z^\star} = \frac{1}{\lambda\,\eta}$$

$$\Rightarrow \quad r^\star + z^\star = \lambda\,\eta\,\phi A \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} L.$$

To get an expression for average quality growth, it is useful to first difference average quality:

$$Q(t) \equiv \int_0^1 q(\nu,t)d\nu \rightarrow \dot{Q}(t) = \int_0^1 \dot{q}(\nu,t)d\nu.$$

Then

$$\frac{\dot{Q}(t)}{Q(t)} = \int_0^1 \frac{\dot{q}(\nu,t)}{Q(t)}d\nu = \int_0^1 \frac{\dot{q}(\nu,t)}{q(\nu,t)} \cdot \frac{q(\nu,t)}{Q(t)}d\nu = \int_0^1 \frac{\dot{q}(\nu,t)}{q(\nu,t)}dv \cdot \underbrace{\int_0^1 \frac{q(\nu,t)}{Q(t)}d\nu}_{=1}$$

where the last equality follows from the independence of quality growth across varieties. By Uhlig (1996), the integral of a continuum of iid random variables with finite variance is equal to the expectation of the variables:

$$\frac{\dot{Q}(t)}{Q(t)} = \int_0^1 \frac{\dot{q}(\nu,t)}{q(\nu,t)}d\nu = E\left[\frac{\dot{q}(\nu,t)}{q(\nu,t)}\right] = (\lambda-1)z^\star.$$

Thus average quality $Q(t)$ grows at a smooth rate despite the stochastic evolution of the individual $q(\nu,t)$s:

$$g^\star = (\lambda-1)z^\star.$$

Combining the three BGP equations yields

$$g^\star = \frac{\lambda\eta\phi A \left(A(1-\phi)/\psi\right)^{\frac{1-\phi}{\phi}} L - \rho}{\sigma + 1/(\lambda-1)}, \tag{13.66}$$

$$r^\star = \rho + \frac{\lambda\eta\phi A \left(A(1-\phi)/\psi\right)^{\frac{1-\phi}{\phi}} L - \rho}{1 + 1/(\sigma(\lambda-1))},$$

$$z^\star = \frac{\lambda\eta\phi A \left(A(1-\phi)/\psi\right)^{\frac{1-\phi}{\phi}} L - \rho}{1 + (\lambda-1)\sigma}.$$

Note that the growth rate is increasing in the innovation step size λ, the R&D efficiency parameter η, and the scale of the workforce L. The growth rate is also decreasing in ρ, σ, and ψ. All of these comparative statics are intuitive. For example, a higher population size increases profits for each variety, thereby encouraging more research and faster growth. There is the same strong scale effect as in the Romer model described above. Just as with

that model, one could entertain diminishing returns in innovation to make growth semi-endogenous, so that sustaining growth requires population growth.

Appendix 13.A.10 provides the social planner's solution to the quality ladder model, which yields the growth rate

$$g^{SP} = \frac{(\lambda - 1)\eta\phi A(1-\phi)^{-1}\left(A/\psi\right)^{\frac{1-\phi}{\phi}} L - \rho}{\sigma}.$$

The social planner (SP) growth rate differs from the decentralized equilibrium (DE) growth rate $g^\star = g^{\text{DE}}$ for three reasons. First, the planner sees only the incremental gain $\lambda - 1$ to innovation, whereas private innovators reap λ. This business stealing is a force for $g^{\text{SP}} < g^{\text{DE}}$. Second, the planner uses intermediates more intensively than intermediate goods monopolists (who discourage downstream purchases by charging a markup over marginal cost). This is a force for $g^{\text{SP}} > g^{\text{DE}}$ because the planner values a new variety more given it will be used more. Third, the planner internalizes the positive knowledge externalities from past to future innovation. Equivalently, the planner sees an innovation as lasting forever in that all future innovations will build upon it. Private profits, in contrast, are truncated by creative destruction. This is a force for $g^{\text{SP}} > g^{\text{DE}}$. To recap, the SP growth rate could be higher or lower than the DE growth rate. That said, calibrations typically find the SP growth rate to be far higher than the DE growth rate. See Jones and Williams (1998, 2000) for early calculations, and Atkeson and Burstein (2019) for a more recent analysis.

13.6 Conclusion

The topic of this chapter—economic growth—is of the utmost importance for welfare. The models and empirical results presented here have shaped our discipline on an unprecedented scale. Virtually all macroeconomic models have a neoclassical core as a backbone irrespective of whether they aim to study business cycles, long-run trends, or topics of international macroeconomics.

The proximate causes of cross-country income differences that follow from a development accounting exercise are crucial in directing research on development economics. The literature has aimed to shed light on the sizable residual called TFP, be it with randomized control trials, the sophisticated exploitation of quasi-natural experiments, or more structural approaches that rely on a combination of theory and (often microeconomic) data.

The literature on endogenous growth is vibrant. The combination of theoretical building blocks outlined in this chapter connect the aggregate growth rate to firm dynamics (entry, exit, and the life-cycle firm growth) and market structure (concentration and market power). Markups play a key role both as a source of misallocation as well as an incentive for entry and innovation. Chapter 22 presents a model analysis in that direction. Recent advances link the literature on growth with the one on industrial organization and derive subtle policy recommendations for promoting research and allocative efficiency. Mapping such theories to available microeconomic data on firms and products fell on particularly fertile ground. This provides exciting research opportunities for the years to come.

Chapter 14

Real business cycles

Kurt Mitman

14.1 Introduction

This chapter introduces the concept of *business cycles* and presents the core framework for business cycle analysis: the real business cycle (RBC) model. This begs the question of what is a business cycle and why the qualification "real" for the model. Business cycles are the fluctuations of the economy around its long-run trend. The model is real, because we will be considering the stochastic version of the growth model developed in Chapter 7, without money or any nominal variables. This chapter can be seen as the natural continuation of Chapter 13, that focused on documenting and accounting for long-run growth patterns. Now, instead will be focused more on understanding short-run fluctuations in the economy.

The term business cycles does not necessarily imply that short-run fluctuations are periodic, like the sine or cosine functions. We begin the chapter by zooming in on the data to examine the short-run movements in aggregates like GDP. We'll discuss ways of extracting the trend from the longer time series and how to interpret the deviations from the trend in the short run. Then we'll document the co-movement of aggregate variables over the "cycle." Next, we'll show how the stochastic growth model can quantitatively be used to rationalize business-cycle fluctuations, the seminal contribution of Kydland and Prescott (1982b). Finally, we'll discuss the successes (and shortcomings) of the RBC model, various extensions, and where the literature has gone since it was introduced.

14.2 Business Cycles: A Historical Overview

The modern usage of the term business cycles refers to the fluctuations in economic activity that an economy experiences over a period of time. We refer to them as cycles because they are characterized by periods of economic expansion followed by periods of contraction or recession. The study of business cycles has its roots in studies of economic crises of the 19th century. The Panic of 1825 was arguably the first economic crisis unrelated to war or another external event, and resulted in a global economic downturn (Bordo 1998). Interest in economic cycles grew as the 19th century saw a recurrence of boom and bust episodes throughout the industrialized world.

Early work was primarily descriptive in nature. Economists documented cyclical relationships in the data that appeared at different relative frequencies. Schumpeter (1939) summarizes the three main types of cycles that had been identified in the data (and named after the economists who identified them) during the late 19th and early 20th centuries:

1. Kitchin Inventory Cycle: A short cycle of about 3-5 years, often related to inventory adjustments.

2. Juglar Fixed Investment Cycle: A medium-term cycle of about 7-11 years, associated with fixed investment.

3. Kondratiev Long Waves: A long-term cycle lasting 45-60 years, driven by fundamental technological innovations.

While the early work did manage to identify cycles and provide evidence of co-movement of different economic variables, there was little distinction between the *impulses* or shocks that are at the origin of fluctuations and the *propagation* of those shocks through the macroeconomy. One of the first to make a distinction between impulse and propagation was Swedish economist Knut Wicksell with his famous metaphor in 1918: "If you hit a rocking horse with a stick, the movement of the horse will be very different from the stick. The hits are the cause of the movement, but the system's own equilibrium laws condition the form of movement." The stick serves as the impulse to the economy, and the rocking in response to the hit the propagation. In the 1930s, economists began to develop theories to explain both the impulses and their propagation throughout the economy.

Schumpeter's theory of business cycles was based on the idea that the impulses were caused by technological innovations developed by entrepreneurs. Innovations should be interpreted broadly here, for example they include the invention of new products and also methods of production. Since coming up with new ideas is an inherently random process, these served as the shocks to the economy. In his theory, Schumpeter argued that innovations tend to occur in clusters, with one breakthrough triggering a cascade of subsequent advances. This clustering of innovations gave rise to the cyclical nature of economic growth. Schumpeter viewed these cycles as fundamental to the capitalist process, with economic development driven by the entrepreneurial introduction of new technologies and methods. This idea of a fundamental link between growth and fluctuations continues to this day.[1]

Schumpeter's theory was, in some sense, "creatively destroyed" before it made a significant impact on economic thought at the time. The Great Depression of the 1930s prompted a rethinking of economic theories that could explain such severe economic downturns. John Maynard Keynes was at the forefront of this intellectual revolution. Keynes argued that insufficient aggregate demand could lead to prolonged periods of high unemployment. He believed that in the face of a decrease in aggregate demand, wages and prices might not

[1]Schumpeter also emphasized the role of credit in facilitating innovations. Entrepreneurs may be constrained financially and have to borrow to finance their innovation activities; see Chapter 19. The banking system plays a crucial role in providing this credit. However, this can also lead to over-expansions of credit, contributing to the cyclical nature of the economy. These ideas are still relevant today: Ben Bernanke would later win a Nobel prize for his work linking the severity of the Great Depression to bank failures and credit disruptions.

adjust quickly enough to restore full employment. As a result, government intervention, in the form of fiscal policy, was necessary to stabilize the economy. One reason, perhaps, why Keynes' theory gained traction was because it both explained the Great Depression and prescribed a way to fight the global slump. His seminal work, "The General Theory of Employment, Interest, and Money" (1936), laid the foundation for what came to be known as Keynesian economics. This school of thought dominated the study of business cycles until 1970s.

In the late 1960s and 1970s, the economic landscape changed with the occurrence of stagflation, a combination of high inflation and high unemployment. This phenomenon was difficult to explain using traditional Keynesian theories that posited a negative relationship between unemployment and inflation. Further, the surge in oil prices following OPEC restrictions on oil supply pointed toward supply-side explanations, which were largely viewed as second-order in the Keynesian theory. At the same time, monetary and fiscal policies in advanced economies during the 1970s appeared ineffective in achieving key policy targets of low and stable inflation. As the Keynesian paradigm struggled to explain stagflation, some economists began to challenge its methodological foundations. Milton Friedman, Edmund Phelps, and Robert Lucas argued that the relationship between inflation and economic activity would change as inflation expectations respond to changes in macroeconomic policy. This type of argument against using reduced-form relatiohsips came to be known as the Lucas (1976) Critique.

In a series of influential papers, Lucas advocated for an approach based on rational expectations—the assumption that individuals understand the structure of the economy and revise their expectations accurately in response to changes in policy. In order to implement such an approach, attention turned to models with microfoundations on the grounds that specifying the primitives of the model in terms of policy-invariant features such as preferences and technologies would avoid the Lucas Critique.

Lucas and Rapping (1969) took a first step towards incorporating rational expectations into business cycle analysis and made significant contributions to the study of labor supply in the context of business cycle fluctuations. Central to Lucas and Rapping's theory is the idea of intertemporal substitution. They posited that individuals make labor supply decisions based on expected future wages. When individuals expect higher wages in the future, they are willing to substitute leisure today for work in the future, and vice versa. In their model, temporary, unanticipated changes in wages lead workers to adjust their labor supply. They argued that observed fluctuations in employment could be largely explained by workers' voluntary decisions to adjust their labor supply in response to unexpected wage changes, rather than by involuntary unemployment. This perspective was in contrast to the traditional Keynesian view, which emphasized the role of demand deficiencies in causing unemployment during recessions. Lucas and Rapping's work suggested that labor market fluctuations might be better understood by focusing on the economy's supply side and workers' intertemporal decisions. Intertemporal substitution has remained central in modern macroeconomic models.

While Lucas emphasized the need for micro-founded macro models, it was Kydland and Prescott who operationalized this vision as a full model of the economy with the development of real business cycle (RBC) theory. Building on rational expectations and intertemporal substitution, they used a stochastic version of the neoclassical growth model, arguing

that business cycle fluctuations result primarily from real shocks—especially productivity shocks—rather than policy interventions. Their work emphasized the importance of understanding the intertemporal decisions of households and firms in analyzing macroeconomic phenomena. From a methodological point of view, Kydland and Prescott essentially introduced into economics the concept of solving models numerically and relying on computer simulations to analyze model behavior (Kydland and Prescott, 2004). By calibrating and simulating dynamic general equilibrium models, they demonstrated that such shocks could replicate observed business cycle patterns. Their approach marked a methodological shift: numerical solution and simulation became central tools of macroeconomic analysis. Prescott viewed the RBC model as a foundational framework for macro, akin to supply and demand in micro. This work established the field of quantitative macroeconomics and became the dominant paradigm for decades.

14.3 A first look at the data

One of the key methodological innovations of the RBC paradigm was to seriously connect a micro-founded business cycle model with real-world data. To begin, we examine the time series of quarterly real U.S. GDP in Figure 14.1, plotted for the post-war period. As emphasized in Chapter 13, U.S. GDP displays a strong upward growth trend. However, in this chapter we will focus on fluctuations around that trend, or the business cycles. Two natural questions arise: how do we define business cycles in the data, and how do we isolate the cycle from the trend component in GDP and other macroeconomic variables?

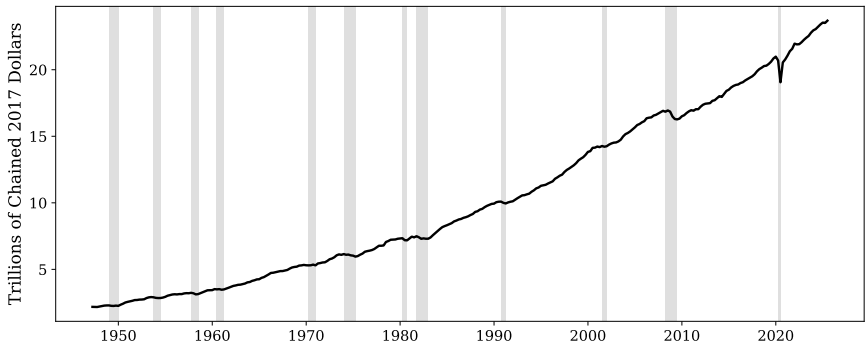


Figure 14.1: U.S. Real GDP and NBER Recessions

In the U.S. the business cycle is defined by its turning points, namely the *peaks* and *troughs*. In between a peak and a trough is a *recession*, and between a trough and peak an *expansion*. Periods of expansion are considered "normal times", whereas recessions are thought of as brief episodes of economic contraction. While the terms seem clear, they are not formal statistical definitions that can be applied to real-time data to determine the current state of the economy. Because of the underlying trend of economic growth, a recession could still be a period of positive GDP growth, albeit at a slower rate. This example highlights the importance of separating the cycle from the trend in the economy. In the U.S., a committee officially determines whether the economy is in a recession. The NBER Business Cycle Dating Committee has officially dated U.S. recessions since 1978. The

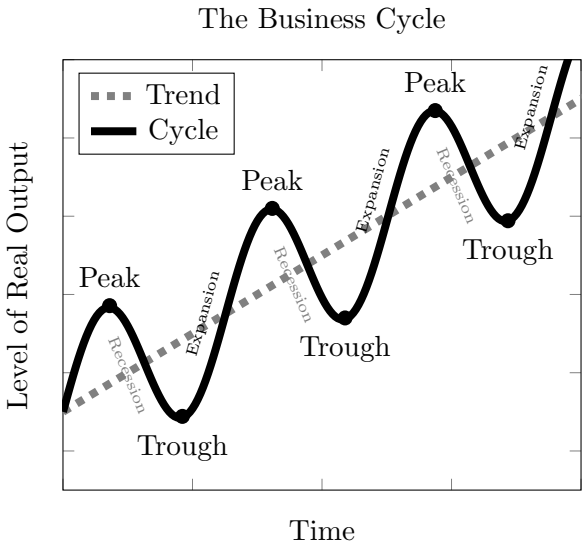


Figure 14.2: Stylized representation of business cycles

committee uses a variety of monthly and quarterly data on measures of economic activity to date business cycles retrospectively. The description of how they determine when the economy is in recession is reproduced below:[2]

NBER Recessions

The NBER's definition emphasizes that a recession involves a significant decline in economic activity that is spread across the economy and lasts more than a few months. In our interpretation of this definition, we treat the three criteria—depth, diffusion, and duration—as somewhat interchangeable. That is, while each criterion needs to be met individually to some degree, extreme conditions revealed by one criterion may partially offset weaker indications from another. For example, in the case of the February 2020 peak in economic activity, the committee concluded that the subsequent drop in activity had been so great and so widely diffused throughout the economy that, even if it proved to be quite brief, the downturn should be classified as a recession. Because a recession must influence the economy broadly and not be confined to one sector, the committee emphasizes economy-wide measures of economic activity.

Given the inherently qualitative nature of the NBER recession dating, we will instead focus on a statistical definition following Lucas (1977) and Kydland and Prescott (1982b) that defines business cycles as recurrent fluctuations of output along a slow-moving trend, and the associated co-movements of other aggregate quantities. The basic idea is that we want to split trending or non-stationary variables, such as GDP, into a non-stationary trend component and a stationary, cyclical component. A stylized representation of the decomposition into trend and stationary is plotted in Figure 14.2. Through the lens of the neoclassical growth model presented in the previous chapter, the goal would be to remove

[2]Source: https://www.nber.org/research/business-cycle-dating, accessed January 5, 2024.

the exogenous deterministic labor-augmenting productivity growth, and transform the model into a stationary one. In the context of the model, that would be easy to achieve. However, identifying the trend in the data presents more of a challenge, since we only observe the combination of trend and stationary components. Further complicating the analysis is that the longer-run trend may not be constant, but could be something more akin to a Kondatiev Long Wave discussed above. To make progress we will discuss techniques for *filtering* the data, that is, extracting trends of different frequencies. In this chapter, we will discuss the most common method for detrending in the RBC literature, the Hodrick-Prescott filter, as well as another commonly used method, the band-pass filter.

14.3.1 Filters

The most common approach to business cycle analysis involves filtering the data to transform it into a stationary series that can be studied in isolation. The idea is to come up with a graph similar to Figure 14.3.

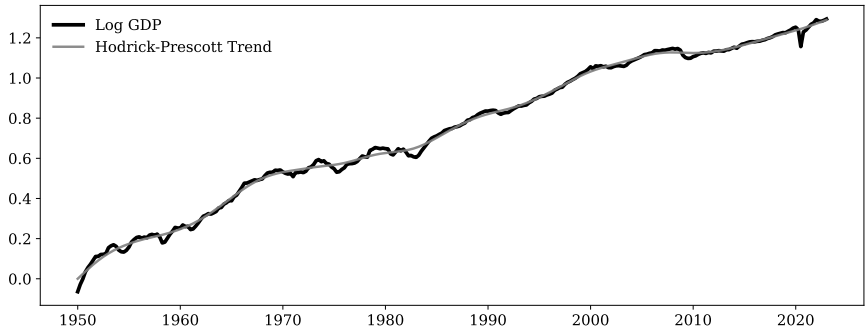


Figure 14.3: Log of U.S. GDP: data and trend from Hodrick-Prescott filter

The optimal way of extracting the trend and cycle components from the data, in principle, should depend on the underlying theory or assumptions about the data generating process (DGP). We can represent the DGP for a macroeconomic time series, Y_t, as a trend-stationary process by assuming that we can decompose them into trend and cycle components, $Y_t = \tau_t + Y_{c,t}$. For example, imagine the trend component is simple exogenous growth $\tau_t = y_0 e^{\gamma t}$ and the cyclical component is simply i.i.d, $Y_{c,t} = \epsilon_t$. One could then simply estimate a linear trend in $\log Y_t$ to capture the trend component, and the residual from that trend would be represent the cyclical one. Another common formulation for the DGP is to assume that the process is difference-stationary. A process that's stationary in first-differences could be represented as $\Delta Y_t = \bar{Y} + \phi(L)\epsilon_t$, where L is the lag operator ($LX_t = X_{t-1}$), $\Delta \equiv (1 - L)$ is the difference operator, and ϕ is a polynomial. For example, a random walk with constant drift could be represented as $\Delta Y_t = \bar{Y} + \epsilon_t$. In finite time series the data can be approximated equally well by difference-stationary and trend-stationary processes (Hamilton, 1994). In practice, therefore, there are many potential DGPs that could be consistent with the time series data. Any filter chosen will be understood to only represent an approximation of the true underlying DGP, and thus will only capture some aspects of the data. Most business cycle analysis therefore does not specify an underlying DGP for the data, but instead proceeds

with filtering in a more informal way. Proceeding in this fashion is appropriate if the results are not sensitive to the particular filter chosen. While this type of robustness analysis was carried out in early RBC contributions (e.g., Prescott, 1986), it is typically omitted in current business cycle analysis. This chapter focuses primarily on the most widely-used filter in business cycle analysis, the Hodrick-Prescott filter.[3]

The Hodrick-Prescott Filter Given a time series $\{Y_t\}_{t=0}^T$, the Hodrick-Prescott (HP) filter decomposes it into a trend component τ_t and a cyclical component $Y_{c,t}$ such that:

$$Y_t = \tau_t + Y_{c,t}$$

The trend component τ_t is obtained by minimizing the following objective function:

$$\min_{\{\tau_t\}} \sum_{t=1}^{T} (Y_t - \tau_t)^2 \tag{14.1}$$

subject to

$$\sum_{t=2}^{T-1} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})\right]^2 \leq \mu,$$

where μ is a parameter that controls the smoothness of the trend component. Smaller values of μ impose smoother paths for the trend component of the time series. In the limit, with $\mu = 0$, $\tau_{t+1} - \tau_t = \tau_t - \tau_{t-1}$ $\forall t$, which imposes a linear trend. Thus, the minimization balances a trade off between tracking the raw data perfectly and following a linear trend.

Under the assumption that the constraint binds, we can re-write the minimization problem using the Lagrangian, where we attached λ as the lagrange multiplier on the constraint:

$$\min_{\{\tau_t\}} \left[\sum_{t=1}^{T} (Y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \left((\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})\right)^2\right]$$

.

In this formulation, μ no longer appears explicitly, and the filter can instead be parameterized by λ, which serves as a smoothing parameter. Higher values of λ impose tighter smoothness constraints on the trend component. In the limit as $\lambda \to \infty$, the extracted trend converges to a linear function. For quarterly data, Prescott (1986) suggested a value of $\lambda = 1{,}600$ as a reasonable choice. However, the appropriate value of λ depends on the frequency of the data. Ravn and Uhlig (2002) show that the HP filter parameter should be scaled by the fourth power of the ratio of observation frequencies. For instance, if $\lambda = 1{,}600$ is used for quarterly data, the corresponding value for annual data would be $1{,}600 \times (1/4)^4 = 6.25$.

To implement the HP filter, one needs to solve the above optimization problem. While there are various methods to do this, a common approach is to use matrix algebra and the system of first-order conditions (FOCs) from the problem. The FOC for τ_t for $2 \leq t \leq T-2$ are given by:

$$\tau_t - \lambda \left(\Delta\tau_t - \Delta\tau_{t-1} - 2\left(\Delta\tau_{t+1} - \Delta\tau_t\right) + \Delta\tau_{t+2} - \Delta\tau_{t+1}\right) = Y_t.$$

[3] See Hamilton (2018) for a discussion of shortcomings of the Hodrick-Prescott filter.

For the end conditions the FOCs will be slightly different (because they cannot rely on data from before $t = 0$ or after $t = T$). But one can show that this system of FOCs has the following form: $(I + \lambda A)\boldsymbol{\tau} = \boldsymbol{Y}$, where I is the identity matrix, $\boldsymbol{\tau}$ is the vector of τ_t, $\boldsymbol{Y}$ is the vector of Y_t, and A is a square pentadiagonal matrix given by:

$$\begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -2 & 5 & -4 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & \cdots & 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

Given a value of λ, we can obtain the trend by solving the system $\boldsymbol{\tau} = (I + \lambda A)^{-1}\boldsymbol{Y}$. Then, we can recover the cyclical component by differencing the data from the trend component. In practice, most common programming languages used by economists (e.g., MATLAB and Python) have built-in functions for HP-filtering the data. There are some things to keep in mind, however, when using the HP filter. First, the choice of λ can influence the resulting decomposition. There is no universally optimal choice for λ, and different applications may require different values (see Marcet and Ravn, 2004, for further discussion). Second, the filter may produce spurious results towards the beginning and end of the sample. Finally, the particular functional form for the HP-filter is not based on economic theory.

Spectrum and band-pass filters In our discussion so far of filtering and separating trends versus cyclical components, we have discussed the idea of trying to isolate cycles at "business cycle frequencies," that is, cycles that occur with a frequency of something like 5-10 years. Ideally, then, we'd like to extract the components of the data at that frequency and remove any longer-run trends that occur with longer frequencies. We typically express time-series data in economics in the time domain, but for stationary time series there exists an equivalent representation in the *frequency domain*, as the sum of sine and cosine functions of various frequencies, amplitudes and phases.[4] Higher frequency means a rapidly changing wave. The sine function has a range of $(-1, 1)$, but the *amplitude* can be adjusted by pre-multiplying the function. The *phase*, or where the sine wave is in its cycle can be adjusted by adding or subtracting from the argument of the function (which shifts the wave to the left or right). Thus, we can write a general sine wave as $B\sin(ft + \phi)$, where B controls the amplitude, f controls the frequency, and ϕ controls the phase.

It is possible to represent any stationary process Y_t as follows:

$$Y_t = \int_0^{\pi} A(\omega)\cos(\omega t)d\omega + \int_0^{\pi} B(\omega)\sin(\omega t)d\omega,$$

[4]To remind you of the properties of sine, the function $\sin(2\pi tf)$ is plotted for difference frequencies f in Appendix Figure 14.A.1. The first sine wave completes a full cycle at $t = 2\pi$, whereas the wave with frequency of 2 completes two full cycles by that time, and the wave of frequency 4 completes four cycles.

where for any ω, $A(\omega)$ and $B(\omega)$ are random variables, and those functions A and B are what define the process. What this expression implies is that the time series is a sum, with random weights, of sine and cosine waves, and the sum is over all frequencies between 0 and π. If we want to isolate a particular range of frequencies, we can imagine a filter that removes any frequency that falls outside of the intended range. Such a filter exists and is called the *band-pass* filter. Using the band pass filter one can isolate and extract cyclical components focusing on any specific range of frequencies (e.g., 5-10 years for business cycles) in the data. A more detailed discussion of the band-pass filter is provided in Appendix 14.A.

14.3.2 Stylized business-cycle facts

Having reviewed tools for extracting business-cycle components, we now present key empirical regularities that motivated the RBC framework. These stylized facts are based on filtered data, primarily using the HP filter, but we also demonstrate their robustness to alternative methods such as first differencing and the band-pass filter. In this section we'll also investigate how the business cycle facts have changed over time, comparing the original statistics from Kydland and Prescott (1982b) to those based on all currently available data. We begin with unconditional moments—specifically, second moments and correlations of aggregate time series. Later, we will examine conditional correlations, following TFP shocks.

One of the defining features of business cycles is the *comovement* of aggregates in booms and recessions. The standard practice in the business-cycle literature is to define co-movements by looking at correlations of aggregates relative to GDP, $Y_{c,t}$. Focusing on contemporaneous correlations, we say that a variable $X_{c,t}$ is:

- *Pro-cyclical* if it comoves positively with output, $corr(Y_{c,t}, X_{c,t}) > 0$.
- *Countercyclical* if it comoves negatively with output, $corr(Y_{c,t}, X_{c,t}) < 0$.

In addition to contemporaneous correlations, to understand propagation it is also helpful to understand which variables *lead* or *lag* movements in output. A variable $X_{c,t}$ is said to be leading output if $corr(Y_{c,t+s}, X_{c,t})$ is highest and positive for some $s > 0$, and is lagging if $corr(Y_{c,t+s}, X_{c,t})$ is highest and positive for some $s < 0$.

Data sources

The data on output, consumption, investment, government consumption, and the price level (here measured as the GDP deflator) come from the National Income and Product Accounts (NIPA) provided by the U.S. Bureau of Economic Analysis (BEA). The data on employment come from the Current Employment Statistics provided by the U.S. Bureau of Labor Statistics (BLS). The data on the unemployment rate is from the Current Population Survey (CPS) provided by the BLS. The data on hours and wages (per hour) are from the non-farm business sector provided by the BLS. The fed funds rate (the policy rate for the central bank) is provided by the Board of Governors of the Federal Reserve System. All quantity variables are expressed in per capita terms by dividing by the civilian non-institutional population aged 16+. All variables are all expressed in logs, except for the federal funds rate. The data are at a quarterly frequency, and all cyclical components are then extracted using an HP

filter with $\lambda = 1{,}600$. We plot the data and HP trend for output, consumption, investment and, hours in Figure 14.4.

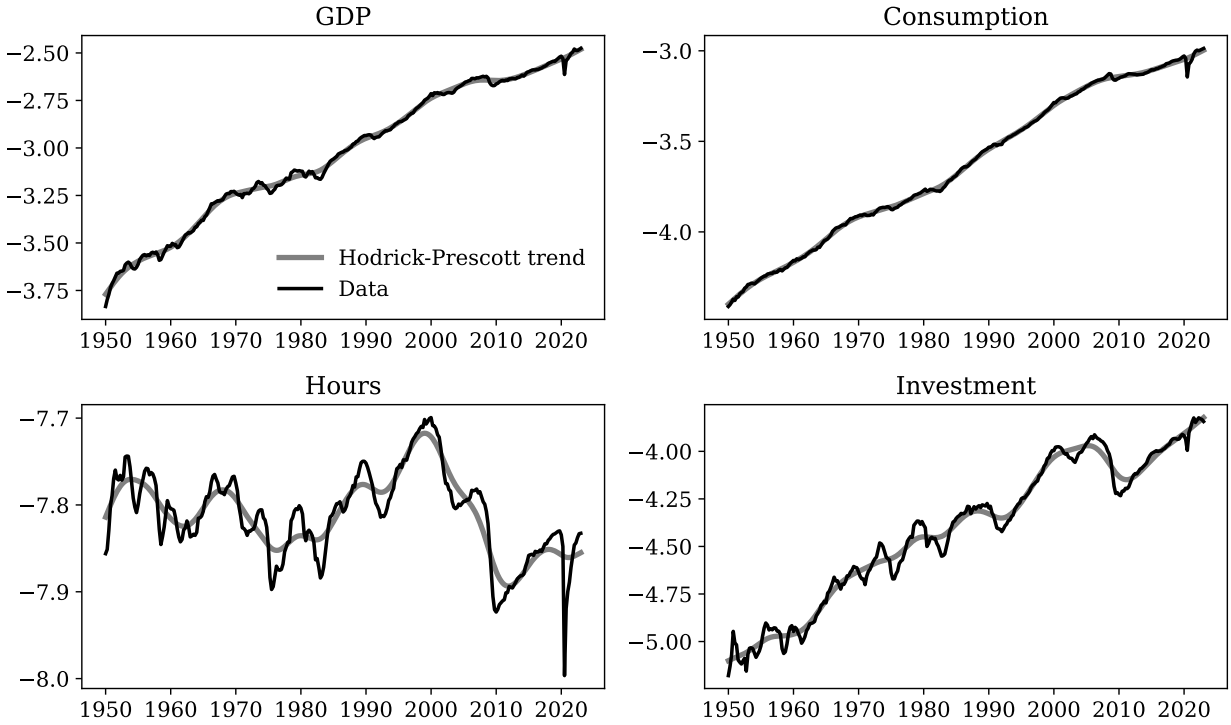


Figure 14.4: Actual and trends of logs of U.S. aggregates

Business cycle facts

The second moments and correlations for the U.S. are listed in Table 14.1 and the main stylized facts are summarized below[5].

Fact 1 **Consumption smoothing:** Consumption is less volatile than output, with a relative standard deviation of 0.65.

Fact 2 **High investment volatility:** Investment is the most volatile component of GDP, with a standard deviation about 2.7 times that of output.

Fact 3 **Labor market dynamics:** Hours worked have similar volatility to output; employment is slightly less volatile, while unemployment is highly volatile and strongly countercyclical.

Fact 4 **Cyclicality:** Consumption, investment, employment, hours, and productivity are all procyclical, while unemployment is strongly countercyclical and lags output.

[5]In Chapter 23, we show that many of these stylized facts hold across countries.

Table 14.1: Business cycle moments U.S. Data 1949-2022

Variable x	Standard Deviation (%)	Relative Std to σ_y	Auto-correlation	Cross correlation of x with $y(t-1)$	$y(t)$	$y(t+1)$
Output (y)	1.63	1.00	0.78	0.78	1.00	0.78
Consumption	1.07	0.65	0.64	0.58	0.75	0.53
Gov. Consumption	3.02	1.85	0.89	0.23	0.15	0.04
Investment	4.37	2.68	0.86	0.61	0.77	0.71
Employment	1.58	0.97	0.81	0.79	0.79	0.48
Hours	2.11	1.29	0.80	0.77	0.86	0.61
Unemployment	15.63	9.57	0.80	-0.79	-0.83	-0.55
Lab. Productivity	1.15	0.70	0.72	-0.03	0.27	0.38
Wages	1.24	0.76	0.72	-0.03	-0.00	0.15
Price Level	0.92	0.56	0.92	-0.02	-0.10	-0.21
TFP	0.90	0.55	0.75	0.17	0.51	0.52
Fed Funds Rate	3.60	2.20	0.97	0.23	0.18	0.08

Fact 5 **Price and wage dynamics:** Despite large swings in quantities, real wages and the price level are much less volatile and less cyclical[6].

Fact 6 **High serial correlation:** All major macroeconomic aggregates exhibit significant serial correlation. Notably, TFP and labor productivity are about as persistent as output itself.

In addition, if one disaggregates production by sector, it is possible to show that most sectors move together over the cycle, though some (e.g., mining) may behave differently. Disaggregating consumption reveals that durable goods are most volatile, followed by non-durables, with services being the least volatile.[7]

The high degree of serial correlation observed in all major macroeconomic aggregates raises an important question: Does this persistence reflect highly persistent shocks to the economy, or do even transitory shocks produce long-lasting effects because of strong internal propagation mechanisms? The fact that TFP and labor productivity are about as persistent as output suggests that models driven primarily by TFP shocks will naturally generate persistent fluctuations in other aggregates—even in the absence of substantial endogenous propagation.

To illustrate business cycle co-movement, Figure 14.5 plots consumption, investment, hours, and TFP as deviations from trend. In each panel, GDP is shown as a dashed line

[6]One of the difficulties in measuring the cyclicality of wages is due to selection. The workers that lose their jobs in recessions are not randomly selected. On average, those workers tend to be lower wage workers. Thus, in recessions the composition of workers shifts towards higher wage workers, mitigating the decline in the average wage.

[7]To show that the stylized business cycle facts are robust to different filtering, in Appendix Tables 14.A.1 and 14.A.2 the analysis is repeated using a band-pass filter and the first difference filter. While the exact numbers clearly are not identical across the different filtering methods, the volatilities and relative volatilities are similar. And, importantly, the cyclicality of all variables (except for wages) is the same across all three filters.

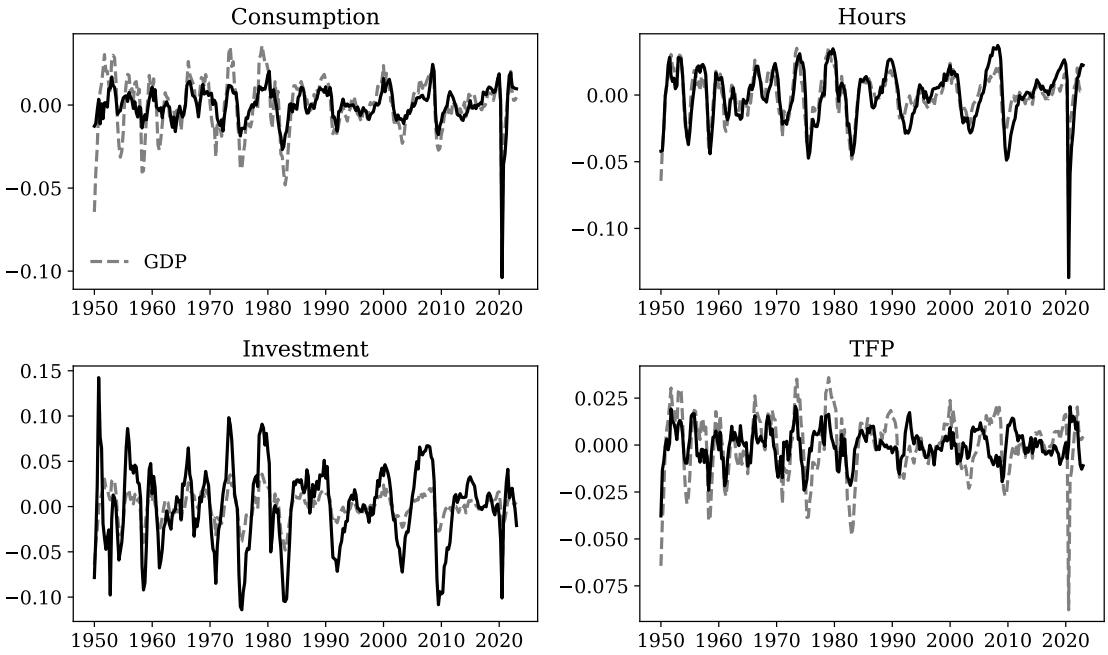


Figure 14.5: Deviations from trend of U.S. aggregates

to highlight both co-movement and relative volatility across series. Notably, the lower right panel shows that, after the mid-1980s, TFP and GDP appear less correlated. To investigate the potential changing nature of business cycles, in Tables 14.2 and 14.3 we recompute the cyclicality using two subsamples, one from 1949-1984 (corresponding roughly to the time period of the original RBC papers), and one for the 1985-2022.

Output is significantly less volatile after the 1980s as compared to before (standard deviation of 1.24% compared to 1.97%), this is sometimes referred to as the "Great Moderation." While the strong co-movement of output, consumption, investment, and labor market variables has been stable over time, some aggregates exhibit significantly different co-movement since the original RBC papers were written. In particular, government consumption switches from being pro-cyclical in the early period to being countercyclical in the second half of the sample. This suggests that fiscal policy has perhaps become more Keynesian over time, using government spending to try to stabilize recessions. The federal funds rate has also become more pro-cyclical, suggesting a change in monetary policy over the time period. Relatedly, the price level goes from being countercyclical to pro-cyclical. Perhaps more surprisingly, wages and labor productivity switch from being strongly pro-cyclical in the early sample to counter-cyclical in the later sample. Indeed, the recessions from 1991 onwards featured so-called "jobless recoveries" in which output recovered quickly, but employment was much slower to recover. Jobless recoveries and countercyclical labor productivity present a challenge for models of productivity-driven business cycles that we will discuss towards the end of the chapter.

Table 14.2: Business cycle moments U.S. Data 1985-2022

Variable x	Standard Deviation (%)	Relative Std to σ_y	Auto-correlation	Cross correlation of x with $y(t-1)$	$y(t)$	$y(t+1)$
Output (y)	1.24	1.00	0.64	0.64	1.00	0.64
Consumption	1.20	0.97	0.56	0.56	0.86	0.43
Gov. Consumption	1.26	1.02	0.83	-0.32	-0.38	-0.39
Investment	3.64	2.94	0.90	0.63	0.79	0.70
Employment	1.56	1.26	0.72	0.69	0.78	0.34
Hours	2.19	1.77	0.74	0.70	0.88	0.52
Unemployment	15.60	12.59	0.74	-0.72	-0.83	-0.43
Lab. Productivity	1.09	0.88	0.76	-0.47	-0.29	-0.05
Wages	1.46	1.18	0.69	-0.21	-0.31	-0.06
Price Level	0.68	0.55	0.93	0.37	0.30	0.14
TFP	0.73	0.59	0.78	-0.33	-0.03	0.10
Fed Funds Rate	2.83	2.29	0.99	0.30	0.29	0.25

14.3.3 Conditional data moments

As discussed at the beginning of the chapter, modern business cycle analysis is interested both in understanding the impulses to the economy and their propagation. A fundamental tool for understanding the propagation of structural shocks is through an *impulse response function* or IRF introduced in Chapter 3.5.2. The IRF traces out the effect of a one-time shock to the economy to current and future values of other aggregates, i.e., the conditional response of variables to the realization of the shock. IRFs, which are linear representations and should be interpreted as approximations applicable to small shocks, visualize the complex interdependencies and dynamic responses in the macroeconomy. Analyzing the IRF is a key tool for understanding the dynamics in the system, and as we'll discuss in the next section we can construct model analogues to IRFs to directly compare model and data. Below, we apply the two most common methods, discussed in Chapter 8, for estimating conditional data moments and IRFs to the RBC context: vector autoregressions (VARs) and local projections (LPs).

Vector autoregression A VAR model for business cycles might be used to forecast labor productivity, output, and investment simultaneously, where each of these variables is expected to affect the others:

$$
\begin{aligned}
A_t &= a_1 + b_{11}A_{t-1} + b_{12}y_{t-1} + b_{13}i_{t-1} + \eta_{1,t} \\
y_t &= a_2 + b_{21}A_{t-1} + b_{22}y_{t-1} + b_{23}i_{t-1} + \eta_{2,t} \\
i_t &= a_3 + b_{31}A_{t-1} + b_{32}y_{t-1} + b_{33}i_{t-1} + \eta_{3,t}.
\end{aligned}
$$

Table 14.3: Business cycle moments U.S. Data 1949-1984

Variable x	Standard Deviation (%)	Relative Std to σ_y	Auto-correlation	Cross correlation of x with $y(t-1)$	$y(t)$	$y(t+1)$
Output (y)	1.97	1.00	0.84	0.84	1.00	0.84
Consumption	0.89	0.45	0.81	0.66	0.75	0.69
Gov. Consumption	4.15	2.11	0.90	0.35	0.26	0.12
Investment	5.00	2.54	0.83	0.59	0.76	0.70
Employment	1.56	0.79	0.90	0.89	0.83	0.59
Hours	2.00	1.02	0.88	0.88	0.90	0.70
Unemployment	15.66	7.96	0.87	-0.89	-0.88	-0.66
Lab. Productivity	1.21	0.61	0.68	0.23	0.63	0.66
Wages	0.93	0.47	0.77	0.13	0.33	0.40
Price Level	1.14	0.58	0.91	-0.17	-0.28	-0.37
TFP	1.05	0.54	0.73	0.42	0.76	0.73
Fed Funds Rate	3.89	1.98	0.95	0.25	0.16	0.01

Defining

$$X_t \equiv \begin{bmatrix} A_t \\ y_t \\ i_t \end{bmatrix}, \boldsymbol{\epsilon}_t \equiv \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \\ \eta_{3,t} \end{bmatrix},$$

we can write the (reduced-form) VAR as:

$$X_t = \boldsymbol{B} X_{t-1} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{B}$ is the matrix of coefficients on the lags (in many empirical implementations more than one lag is used) and $\boldsymbol{\eta}_t$ is the vector of innovations. As discussed in Chapter 8, the key identification issue is what we will assume about the relationships among contemporaneous ηs. To implement theoretical restrictions, we first write the (structural) VAR as

$$\boldsymbol{A} X_t = \boldsymbol{F} X_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{A}$ now is a matrix that captures the contemporaneous relationships among the variables in X_t and ϵ_t is a vector of structural shocks (as opposed to just reduced-form innovations). Let us, as also proposed as an example in Chapter 8, use timing restrictions:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Thus, we assume that the first variable (TFP) does not respond to contemporaneous movements in output (since the off diagonal elements in the first row are zero). The second variable, output, responds to productivity, but not to investment. Finally, investment can respond contemporaneously to both productivity and output. To solve the model, we invert the A matrix, then estimate the reduced-form VAR, and then back out the structural shocks

from the reduced-form errors.[8] The impulse response of variable i at time $t+s$ to a shock to variable j at time t can then be backed out by solving for $\partial X_{i,t+s}/\partial \epsilon_{j,y}$.

The identifying restrictions used here align with the RBC model presented in the next section: if TFP shocks drive business cycles and capital is predetermined, it is reasonable to assume that productivity does not immediately respond to output or investment. Likewise, since capital investment takes a period to become productive, output and productivity should not respond to current investment. Figure 14.6 shows the impulse responses to a TFP shock based on an SVAR estimated using U.S. data. The IRFs show significant co-movement between productivity, output, and investment from the shock to TFP. These IRFs suggest that shocks to productivity are promising candidates for explaining the unconditional moments described in the preceding section. Productivity shocks could generate the right co-movements and also the relative volatilities between output and investment.

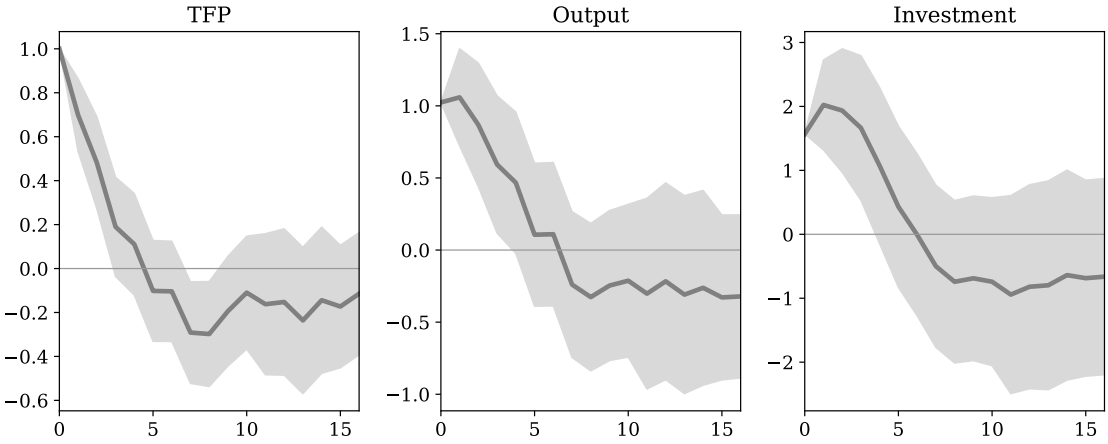


Figure 14.6: Impulse responses to TFP shock estimated with structural VAR

Local projections In the VAR above, we used timing restrictions to identify TFP shocks. If we assume that our measure of TFP (calculated as the Solow residual) is an exogenous shock, we can estimate IRFs to these shocks by means of local projections. The results are plotted in Figure 14.7. The impulse responses are qualitatively similar to those derived from the SVAR, which is reassuring and further reinforces the notion that TFP shocks are a potential important driver of business cycle fluctuations. We now turn to the theory behind the real business cycle model.

14.4 Real business cycle model

To develop a micro-founded framework for business cycle analysis, Kydland and Prescott extended the neoclassical growth model from Chapter 4 by introducing stochastic TFP, giving rise to the real business cycle (RBC) model outlined in Chapter 7. This stochastic neoclassical growth model is the foundation of modern business cycle theory and serves as

[8] See the box on recursive identification in Chapter 8.

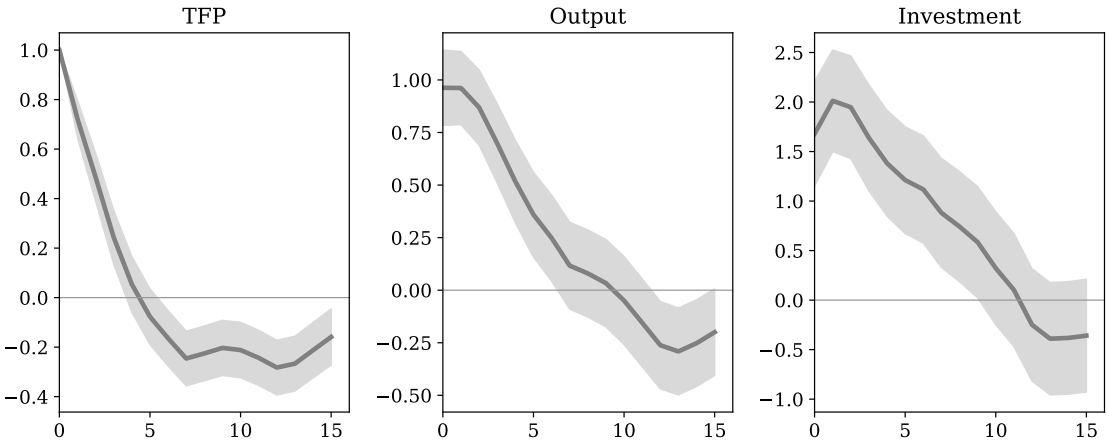


Figure 14.7: Impulse responses to TFP shock estimated with local projection

the core for both New Keynesian (Chapter 18) and International Macro models (Chapter 23). Before presenting the full model with capital, we begin with a simplified version that includes only labor supply. This version admits a closed-form solution and serves to illustrate why labor alone cannot account for the key business cycle facts.

14.4.1 Simple business cycle model

Consider a representative household that maximizes its expected discounted utility. Time is discrete and lasts forever. The utility function is characterized by log preferences that are separable between consumption c_t and leisure $1 - \ell_t$, where ℓ_t represents labor supply. The utility function is given by:

$$U = E_0 \sum_{t=0}^{\infty} \beta^t \Big(\ln(c_t) + \phi \ln(1 - \ell_t) \Big). \tag{14.2}$$

The production function of the representative firm is linear in labor and subject to TFP shocks z_t:

$$Y_t = z_t \ell_t, \tag{14.3}$$

where we assume that z_t follows an AR(1) process in logs, $\log z_t = \rho \log z_{t-1} + \sigma\epsilon$. Perfect competition and linear production on the firm side imply that the equilibrium wage will be equal to TFP, $w_t = z_t$.

The representative household receives income from supplying labor and has access to trade a risk-free one period bond that is in zero net supply:

$$c_t + B_t = w_t \ell_t + R_t B_{t-1}, \tag{14.4}$$

where R_t is the gross real rate and B_t are the bonds purchased by the household in period t.

Equilibrium and labor supply In equilibrium, the household's choice of labor supply balances the marginal utility of consumption against the disutility of labor. Substituting out for consumption using the budget constraint, and taking the first-order condition with respect to ℓ_t, we arrive at the static intratemporal optimization condition:

$$\frac{w_t}{c_t} = \frac{\phi}{1-\ell_t}. \tag{14.5}$$

We can derive the Euler equation from the intertemporal optimality condition:

$$\frac{1}{c_t} = \beta \mathbb{E}_t \frac{R_{t+1}}{c_{t+1}}. \tag{14.6}$$

Constant labor supply is optimal in equilibrium The last step in characterizing the equilibrium is to recognize that since the bond is in zero net supply and we have a representative household, it must be the case that in equilibrium the household holds no bonds. Next, we can eliminate consumption from the intratemporal labor supply condition using the budget constraint and setting $B_t = 0$ for all t to arrive at:

$$\frac{\phi}{1-\ell_t} = \frac{w_t}{w_t \ell_t} = \frac{1}{\ell_t}. \tag{14.7}$$

In this setup, the equilibrium labor supply is constant across periods, independent of the wage, and determined solely by preferences. Although the economy experiences fluctuations, there is no amplification or propagation, only immediate responses to shocks. Since output equals fixed labor input times TFP, output, consumption, and TFP are perfectly correlated, while hours worked remain unchanged. This is clearly inconsistent with the business cycle facts presented earlier in the chapter. At first glance, this result may seem puzzling. As shown in Chapter 12, log-log preferences imply a Frisch elasticity of $e_F = (1-\ell)/\ell$. If households work one-third of their available time, this gives a Frisch elasticity of 2—suggesting highly elastic labor supply. So why is there no response of hours to fluctuations in the real wage? The key lies in the definition of the Frisch elasticity: it measures labor supply responses holding the marginal utility of consumption constant. In this model, households cannot transfer resources across time, as there are no savings or borrowing instruments. Thus, the real interest rate must adjust to make it optimal for households to choose not to intertemporally substitute across time. When TFP is high, consumption rises and the real interest rate falls to discourage saving. If the real interest rate were constant (or if households could save) they would increase the labor supply when wages are high and smooth consumption using savings. As discussed in Chapter 13, RBC models typically assume balanced-growth-consistent preferences, which imply constant labor supply along the growth path. Higher TFP has both income and substitution effects, but the assumption of linear production implies that they both cancel out in equilibrium. Constant labor supply is thus a consequence of balanced-growth preferences and consumption and output that are perfectly correlated.

Keep these results in mind when reading Chapter 18 on the New Keynesian model. The production side of that economy is identical to the one presented here. However, the introduction of nominal rigidities gives monetary policy the ability to control the real interest rate, which drives fluctuations in consumption. Wages adjust to clear the labor market.

Unlike the RBC model, where TFP shocks are the primary driver, here fluctuations in the real interest rate become the key source of output variation, operating through the intertemporal substitution in consumption channel.

To sum up the simple model, introducing TFP shocks and elastic labor supply is not sufficient for matching the facts on business cycles. As we will see next, with Cobb-Douglas production and capital that depreciates at a rate consistent with the data, we break the constant labor supply assumption, and generate volatile hours and investment over the cycle.

14.4.2 Core RBC model

We now move to describe the more general version of the model, referred to as the core RBC model. We will first discuss its main features and strengths, then discuss extensions developed to address some of its shortcomings. Given the thorough treatment of the neoclassical growth model in the Chapter 7, our exposition will be rather brief. In particular, we focus on the planning problem, relying on the equilibrium characterization presented in Section 7.5 of that chapter. For simplicity, we abstract from population and productivity growth; alternatively, we could explicitly include growth and then transform the model into a stationary one.[9]

Preferences The economy is populated by a measure one identical households endowed with 1 unit of time that they can allocate to work or leisure. They have preferences over consumption c and leisure $1-\ell$ represented by the period utility function $U(c, 1-\ell)$. We assume that households maximize expected discounted utility, and that their preferences thus can be written as

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(c_t, 1-\ell_t).$$

Technology Output is produced according to a production function F that takes capital k_t and labor ℓ_t as inputs:

$$Y_t = z_t F(k_t, \ell_t),$$

where z_t is stochastic TFP. F is assumed to have the usual properties, i.e., it has constant returns to scale, it is concave in both arguments, and satisfies the Inada conditions. We assume that log productivity follows an AR(1) process, $\log(z_t) = \rho_z \log(z_{t-1}) + \sigma_z \epsilon_t$, where $\epsilon_t \sim N(0,1)$ are iid random variables. Capital depreciates at constant rate δ and it takes "time to build." That is, investment in period t becomes productive in period $t+1$. The evolution of the capital stock can be written as:

$$k_{t+1} = (1-\delta)k_t + i_t.$$

Virtually all of the analysis for the case of the neoclasical growth model goes through nearly unchanged for the RBC model. In particular, one could show that the competitive

[9] As discussed earlier, balanced growth preferences are homothetic, so renormalizing the model under constant growth rates is relatively trivial.

equilibrium is efficient (i.e., the first welfare theorem holds) and that we can characterize the competitive equilibrium allocations using the solution to the social planner problem. The main difference is that allocations now depend on the history of realizations of the productivity shock. Thus, we proceed directly to analyzing the social planner problem in recursive form.

The planning problem for this economy can be written recursively as:

$$V(k,z) = \max_{k',\ell} U(zF(k,\ell) + (1-\delta)k - k', 1-\ell) + \mathbb{E}\left[V(k',z')|z\right],$$

where we've substituted out for consumption using the budget constraint. The solution to this problem delivers decision rules for capital accumulation, $k' = g(k,z)$, and one for hours worked, $\ell = h(k,z)$.

The first-order conditions for the maximization problem (assuming differentiability of the value function) are given by

$$zF_2(k,\ell) = \frac{U_2(c,1-\ell)}{U_1(c,1-\ell)} \tag{14.8}$$

and

$$U_1(c,1-\ell) = \beta\mathbb{E}[V_1(k',z')|z], \tag{14.9}$$

where U_1 is the marginal utility of consumption, U_2 is the marginal utility of leisure and V_1 is the first derivative of the value function with respect to capital (its first argument). The envelope condition reads as

$$V_1(k,z) = \Big(zF_1(k,\ell) + 1 - \delta\Big)U_1(c,1-\ell). \tag{14.10}$$

Using this in equation (14.9) we obtain

$$U_1(c,1-\ell) = \beta\mathbb{E}\left[(z'F_1(k',\ell') + 1 - \delta)\,U_1(c',1-\ell')|z\right]. \tag{14.11}$$

The key optimality conditions of the real business cycle model are (14.8) and (14.11). The intratemporal optimality condition (Equation (14.8)), equates the marginal product of labor (what would be the wage in the competitive equilibrium) with the marginal rate of substitution between consumption and leisure. Temporary positive shocks to z lead to increases in labor supply, as households intertemporally substitute leisure away from a time when the marginal product of working hours is high. The strength of this response depends on the Frisch elasticity, discussed in Chapter 12.

Equation (14.11) is the standard intertemporal Euler equation, which equates the marginal utility of consumption today to the expected marginal utility of consumption tomorrow, adjusted by the time discount factor β and the stochastic rate of return on capital, $z'F_1(k',\ell') + 1 - \delta$, which in turn equals the gross real interest rate in the competitive equilibrium. To the extent that shocks to productivity are persistent, a positive shock to productivity today raises productivity tomorrow. That in turn, *ceteris paribus*, increases the marginal product of capital, which induces households to reduce consumption and increase investment tomorrow. The strength of this response depends on the intertemporal elasticity of substitution discussed in Chapters and 4 and 12.

Despite the RBC model being a fundamentally non-linear one, the original work by Kydland and Prescott (and much of the RBC literature that followed), solved the model by linearizing it around the non-stochastic steady state.[10] The linearization method is discussed in detail in Chapter 10. As noted in Section 10.6 of that chapter, linearizing the model yields linear decision rules for the aggregate variables as functions of the two states: capital k, and TFP, z, for example $y = G(k, z)$ and $k' = H(k, z)$. Using these functions, we can construct impulse response functions for the model by drawing a shock $z_0 = z_s s + \epsilon_0$ at the steady state, and then progressively applying the policy function (assuming no further shocks). This generates a sequence $y_0 = G(k_s s, z_0)$, $y_1 = G(H(k_s s, z_0), z_1)$, $y_1 = G(H(H(k_s s, z_0), z_1), z_2)$, and so on, where we've substituted the dynamics for k using H and $z_t = \rho z_{t-1}$. The sequence $\{y_0/\epsilon_0, y_1/\epsilon_0, \cdots\}$ would represent the model IRF to a one-time TFP shock. The linearized functions are essentially equivalent to a VAR representation: since the composition of linear functions is linear, we can represent a variable at time t as a linear function of lags of the variable. The model-implied impulse response can then be compared to its empirical counterpart—estimated using a VAR or local projection—to assess how well the model captures the propagation dynamics observed in the data.

IRF estimation with MIT-shocks

Analogous to how LPs and VARs estimate the same IRFs, Boppart, Krusell, and Mitman (2018) recently have shown how models can be linearized by computing the perfect-foresight impulse response of the economy to an unexpected shock—a so-called "MIT shock."[a] Instead of having a recursive linear representation of the aggregate variables as in the example above, the model is linearized in *sequence space*, where the only state variable is the time since the shock. The model can then be simulated simply by drawing shocks and superimposing them.[b] In the case of the representative-agent model, this is not necessarily advantageous numerically, since the recursive formulation only has two state variables, whereas for the impulse responses we need to keep track of t values for the impulse response. But for heterogeneous-agent models, discussed in Chapter 21, the recursive formulation has an infinite-dimensional state variable (the wealth distribution), and so linearizing the model in the sequence space a la Boppart et al. (2018) is more efficient computationally and conceptually easier to implement.

[a] Not so named because the author of this chapter is the MIT-man, but coined by Sargent (a Harvard alumnus) to criticize that type of analysis that was currently being carried out at MIT as being inconsistent with rational expectations. But, as Boppart et al. (2018) showed, if the model is linear, then solving for the "MIT shock" is fully consistent with the rational expectations equilibrium.

[b] Auclert and Mitman (2018) show how to estimate and simulate models using simulated method of moments in the sequence space.

14.4.3 The Kydland-Prescott blueprint for business-cycle analysis

In addition to providing a framework for business-cycle analysis, Kydland and Prescott (1982b) essentially created a blueprint for how modern, quantitative macroeconomics would

[10] Subsequent work showed that the RBC model is well approximated by linear approximations because the variance of TFP shocks is small.

be conducted. The starting point for any paper is a precise economic question. Questions can either be positive (e.g., "Can technology shocks account for the co-movement in output, consumption and investment in the data?") or normative (e.g., "How should we optimally provide unemployment insurance over the business-cycle?") in nature. Importantly, questions are about measurement and answers are numbers. The key ingredient to answering the question is a structural theory of the economy (a model). The model is a measurement device used to derive the quantitative implications of the theory. The model should be chosen from a set of "well-tested" theories, i.e., those that are based on numerous micro-econometric studies. The model can be extended to include the necessary ingredients or frictions to answer the question. Next, the parameters of the model need to be chosen, or the model needs to be "calibrated." The original blueprint of Kydland and Prescott (1982b) was to calibrate the model along some dimensions of the data and then used to explain other dimensions of the data. They calibrated the model to be consistent with "long-run" facts (e.g., growth) and then evaluated it based on how well it explained business cycle facts. Finally, the model is solved numerically on the computer to solve for the equilibrium process and run the computation experiment that answers the original economic question. For example, the answer to Kydland and Prescott's original question was that shocks to TFP could explain roughly 70% of the fluctuations in output over the business cycle.

Of the steps outlined above, all should be familiar, except perhaps the "calibration" step, which we revisit here in the context of the RBC model. Calibration was first introduced in Chapter 8, where it was applied to a simpler, deterministic environment. In this chapter, we apply the same conceptual framework to a richer dynamic setting. The general strategy is to use empirical moments at the macro level, alongside microeconomic evidence from households or firms, to discipline the model's preference and technology parameters. In Prescott's view, the purpose of calibration is not to maximize statistical fit, but rather to choose parameters that align with established economic theory and match key stylized facts. The aim is to construct a model consistent with some dimensions of the data and assess how well it explains others. Calibration also made it feasible to work with models that were, at the time, too complex for structural estimation, enabling their use in quantitative analysis.

Calibration and measurement The basic question of Kydland and Prescott (1982b) was how much fluctuations in de-trended macro aggregates could be explained by "technology shocks." The idea behind the shock was one to aggregate TFP, as described in the RBC theory above. The question did not preclude other shocks from being important for business cycle fluctuations (like government-spending shocks, or shocks to people's preferences), but focused on analyzing the role of one particular shock. The question arises, however, how does one discipline the shock process to be fed into the model?

The starting point for measuring TFP was Solow's growth accounting approach, discussed in several different contexts earlier in the book (e.g., in Chapters 2 and 13), thus using the assumption of a CRS production function, perfect competition and measurement of inputs and output, and profit maximization. Measurement of the capital stock is not immediate: the national income and product accounts typically only include measures of investment and depreciation, but not actual measures of the capital stock. The stock of capital is usually inferred from a "perpetual inventory" method. The basic idea is to make an initial guess for

the capital stock, then based on measures of investment and estimates of the depreciation rate of capital δ, compute forward the time series of capital using the accumulation equation from the growth model: $k_t = (1-\delta)k_{t-1} + i_t$. A typical value estimated for δ at a quarterly frequency is 0.025.

Direct use of Solow's method involves using time-series data on capital and labor shares. Since these had not varied greatly over the postwar period (up until the mid-1980s at least), a Solow residual series based on computing the average share and then applying this share period by period did not generate large errors. We set $\alpha = 0.36$ so that it approximates the value of the capital share. Then, the TFP series z_t can be measured as

$$\log z_t = \log y_t - \alpha \log k_t - (1-\alpha) \log \ell_t.$$

Fluctuations in output at business cycle frequencies can arise from changes in inputs or from changes in TFP. By computing the Solow residual, we can gauge the extent to which technology shocks drive to business cycle fluctuations. We can also estimate a stochastic process for those shocks to productivity to use as an input to the RBC model. A result is that the process for z_t is well approximated by an AR(1) process. Typical estimates on post-war data find that the persistence is quite high ($\rho_z = 0.95$ at a quarterly frequency) with a standard deviation of innovations of $\sigma_z = 0.007$.

Limitations of the approach

One of the main criticisms of this approach to calibration is that technology shocks are not measured directly; instead, they could simply be errors in the measurements of inputs and outputs (see the discussion of hard-to-observe utilization rates below). Negative shocks are challenging to understand too. On the other hand, our technologies do advance over time and the notion that these advances are not perfectly even—occur at faster and slower rates over time—is quite natural. A period of slow technological development then implies negative shocks but only relative to trend.

A related point is that multi-sectoral versions of the RBC model must have technology shocks that are highly positively correlated across sectors, since sectoral outputs (and employment) have this feature. Is such an assumption reasonable: why would the construction sector and the service sector experience positive technology shocks at the same time? Clearly, there are general technological developments—such as IT—that benefit all sectors, but it is far from clear that these dominate.

Returning to the possibility of mismeasurement, consider variable capacity utilization as well as quality improvements in the capital stock. During booms, firms would tend to use their capital more intensively, leading to higher measured TFP. Conversely, during recessions, the utilization of inputs drops, leading to lower TFP. To correct for this we can adjust capital K_t for its utilization rate u_t: $K_t^* = u_t K_t$. The composition of labor and capital can change over time. For instance, an experienced workforce or newer machinery can lead to higher output even with the same quantity of inputs. Thus, methods have been developed to adjust labor ℓ_t for changes in workforce composition, education, and experience, and to adjust capital K_t for changes in the quality and type of capital goods.

Before seeing the model in action in the next section, it's important to note that the research program undertaken by Kydland and Prescott was not designed or intended to explain business cycle fluctuations solely with technology shocks. The question was simply: how much could technology shocks account for fluctuations? Their original agenda included as a next step to introduce monetary features into the model and evaluate how much they contribute to fluctuations. The success of technology shocks in explaining the lion's share of business-cycle fluctuations is what ultimately led the researchers to instead increase the complexity and richness of the real setup. Given our previous discussions of the changing cyclicality of productivity post-1985, had Kydland and Prescott developed business-cycle theory in the 2000s it's quite possible that they would have continued to explore the importance of monetary (and other) phenomena.

14.4.4 RBC model in action

After having estimated a process for TFP and specified the production technology, the final step of the calibration is to specify and parameterize household preferences. Here, again, Prescott wanted to restrict the degrees of freedom to be consistent with economic theory and empirical evidence. Given that the RBC framework was built on the neo-classical growth model, the preferences should be consistent with balanced growth. That restricted preferences over consumption to be of the CRRA form. Later, King et al. (1988) proved that preferences over consumption and leisure that were consistent with balanced growth and constant hours had to be of the form:

$$\frac{(c_t v(1-\ell_t))^{1-\sigma}-1}{1-\sigma}.$$

Here, we take v to bea strictly increasing, strictly concave function, satisfying $v'(0)=\infty$.[11] In the original calibration, Prescott argued that balanced growth with increasing wages and consumption implied a unitary elasticity of substitution between consumption and leisure, and thus specified the aggregator between consumption and leisure as Cobb-Douglas, $c^{1-\phi}(1-\ell_t)^{\phi}$. The question is then how to parameterize the shares, ϕ, the CRRA parameter, σ, and the discount factor, β. Again he turned to the implications of the growth model. In steady state the real interest rate is equal to the growth rate of consumption times σ divided by β. Then given data on the average real interest rate and consumption growth, β could be backed out as a function of σ. The value for σ was chosen based on other estimates that tried to measure it in the data based on life-cycle consumption patterns and the response of consumption and stock portfolios. These data suggested that $\sigma=1$ was a reasonable value, implying log preferences. This then implies a quarterly value for $\beta=0.99$, representing an annual 4% discount rate. The period utility function was then given by:

$$u(c_t, 1-\ell_t) = (1-\phi)\log c_t + \phi \log(1-\ell_t).$$

The final parameter to be calibrated is ϕ. Given that the time endowment is normalized to 1, there are 24 hours in the day and people typically work 8 hours, we can pick the value

[11] Boppart and Krusell (2020) later generalized these preferences to accommodate balanced growth and falling hours, consistent with recent macro data across the OECD, or hours in the U.S. from a longer-run perspective.

of ϕ to generate $l_t = 2/3$. Using the first-order conditions of the model equation (14.8) and multiplying both sides by ℓ and dividing by c, with our functional forms we arrive at:

$$\frac{\phi}{1-\phi}\frac{\ell_t}{1-\ell_t} = (1-\alpha)\frac{y_t}{c_t}.$$

With α specified, and the average of y_t/c_t measurable in the data, given a target of $\ell_t = 1/3$ we can then solve for $\phi = 0.6325$.

Now that we've calibrated the model, we can solve for the dynamic system and simulate the model time series to see how they compare to U.S. business cycle statistics. Section 10.6 in Chapter 10 describes the methodology used to solve the model using linearization techniques, with Example 5 detailing a specific application to the calibrated RBC model used in this chapter. The second moments are collected in Table 14.4. All variables are treated as in the data, as long deviations from an HP filtered trend with smoothing parameter 1,600.

Table 14.4: Business cycle moments in core RBC model

Variable x	Standard Deviation	Relative Std. Dev. σ_y	Auto-Correlation	Cross correlation of x with y
Output y	1.33	1	0.73	1
Consumption c	0.42	0.32	0.81	0.90
Investment i	4.13	3.10	0.72	0.99
Hours n	0.64	0.48	0.72	0.98
TFP z	0.92	0.69	0.43	0.99

The key observation from the table is that TFP shocks successfully generate business cycle fluctuations: output fluctuates significantly in response to technology shocks. The volatility of output is larger than that of TFP, implying that the model generates amplification. Consumption, investment, and hours are all pro-cyclical, as in the data. Investment is more volatile than output, while consumption is less volatile than output, both relationships consistent with the empirical findings.

While the model is successful in generating fluctuations in output, the volatility in the model is lower than that in the data (about 1.33% in the model, compared to 2% in the 1949-1984 time period). Hours in the model are less volatile than output, whereas in the data they exhibit approximately the same volatility. This suggests that the Frisch elasticity of 2 is actually *too low* to match the fluctuations in hours worked over the cycle (which in turn would help increase volatility in output). Finally, the correlation of all variables with output is higher than in the data. This is likely due to the fact that we are only considering a single impulse to the model (the TFP shock) which generates the procyclical co-movement between all variables. In reality, the economy likely experiences many kinds of shocks (we will discuss government spending shocks later in this chapter) that exhibit different co-movement patterns. Simulating the model with additional shocks could bring down the strong correlations.

The results from the core RBC model can be understood by analyzing the two key forces in the model. One is how willing households are to work more when the returns to working—measured by the marginal product of labor (MPL)—rise. The second force involves

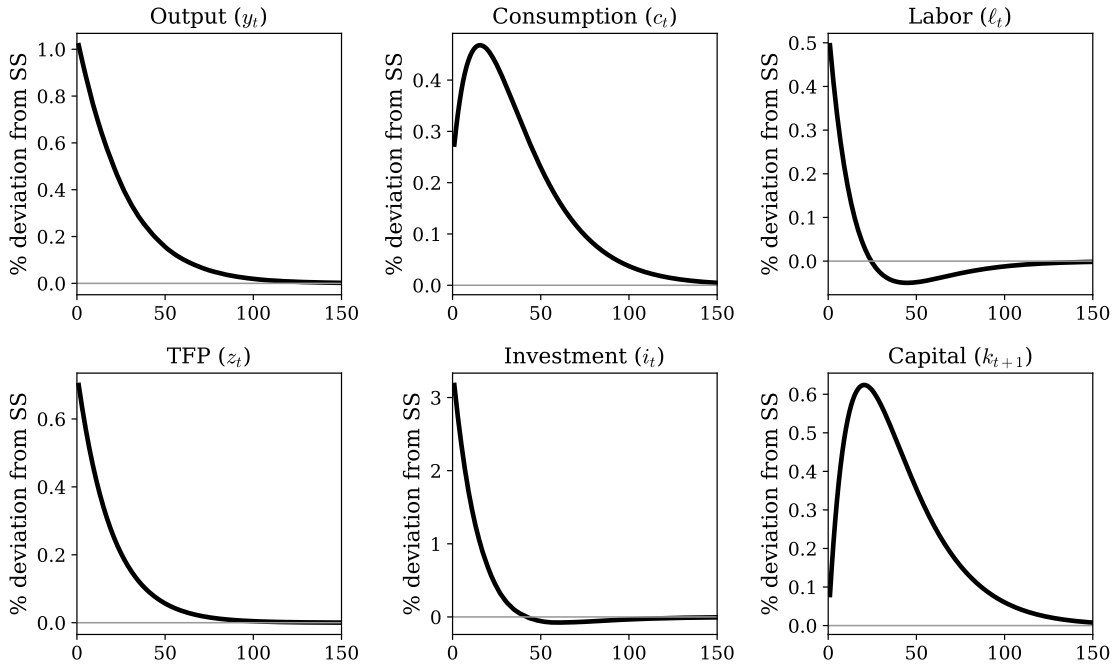


Figure 14.8: TFP shock in RBC model

the willingness of households to smooth consumption over time in the face of fluctuating income. When a positive productivity shock occurs, the incentives to work increase (MPL rises) and output increases (about 1%). In the simple model without capital, consumption rose one-for-one with output, leaving labor supply unchanged. With capital accumulation, however, households can smooth consumption by adjusting savings. As a result, consumption increases by less than output (about 0.3%), while investment responds strongly (over 3%). Because consumption rises only modestly, the response of labor supply is more closely tied to the Frisch elasticity (here, hours increase by about 0.5% because consumption does increase). One way to think about the RBC model is that labor demand is fluctuating (as MPL moves) and moves along the labor supply curve, which in this case is roughly a line with slope equal to the Frisch elasticity. The adjustment is possible because capital accumulation allows households to smooth consumption. This highlights the importance of capital for the RBC model to generate amplification and business cycle co-movement as in the data.

To better explain the mechanism in the model in response to a TFP shock, we can plot the impulse response functions in Figure 14.8. The IRFs show us in response to the impulse (the TFP shock) what happens to all of the endogenous variables in the model. By comparing how much output increases relative to TFP, we can get a measure of the amount of amplification in the model. Similarly, we can measure the propagation of shocks by looking at how long the increase in output lasts relative to the persistence of the exogenous process of TFP. Finally, the IRFs show us the shape of the responses, where they look more linear or hump-shaped. The basic RBC model generates amplification, but does not generate much actual propagation of TFP shocks. We can illustrate this more concretely by solving the model with less persistent processes for TFP. Figure 14.9 plots the impulse responses for the baseline calibration ($\rho = 0.95$) and also for iid TFP shocks ($\rho = 0$) and TFP shocks with a

half-life of one quarter ($\rho = 0.5$). The persistence of output is essentially the same as the TFP process across the three impulse responses. The less persistent processes generate more amplification on impact (with hours and output responding by more), because households are more willing to intertemporally substitute their labor supply. Households respond to the temporary rise in wages by increasing labor supply, but rather than immediately consuming the additional income, they primarily save it through higher investment. This allows them to smooth consumption over time.

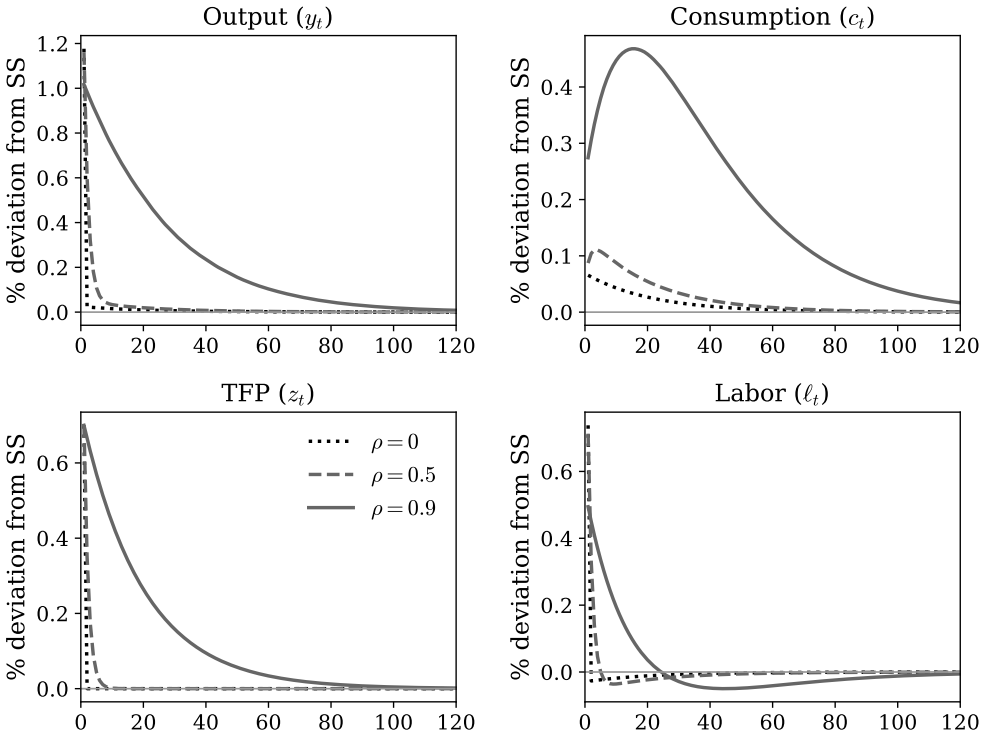


Figure 14.9: Sensitivity to TFP shock persistence in the RBC model

To further illustrate the sensitivity of the model to the two key forces, we explore sensitivity to the intertemporal elasticity of substitution (recall with CRRA preferences the EIS is $1/\sigma$) and to the Frisch elasticity. In the baseline model we had an EIS of 1 and a Frisch elasticity of 2. We can recalibrate the model with higher and lower Frisch elasticities (0.5 and 10), and higher and lower EIS (0.5 and 2), and compare the IRFs to those in the baseline model. In Figure 14.10 we plot the IRFs for those alternate calibrations.

Focusing first on the left panel, if we increase the Frisch elasticity, households are more willing to intertemporally substitute labor supply, and we get more amplification. Hours increase by more in response to the TFP shock, which leads to higher output, and more volatile investment as households save the extra earnings to smooth consumption going forward. With a lower Frisch elasticity we get the opposite, more dampening. The core RBC model gave a response of hours and output that was too low relative to the data, suggesting that a calibration with a higher Frisch elasticity may improve the fit. However, at the same time that leads to more volatile investment, which was already too high relative to the data.

Turnings to the right panel of Figure 14.10, a higher intertemporal elasticity of substitution also leads to more amplification of output and labor supply in response to TFP shocks. On impact, however, the response of consumption, is non-monotonic in the EIS (it actually falls in both cases). When households value consumption smoothing less, they are more willing to intertemporally substitute to take advantage of the productivity shock. Households are willing to increase consumption by less today and lower leisure more to increase investment to take advantage of the persistently higher TFP. That leads to a bigger boom in consumption in the medium run, after houseboulds have built up the capital stock. Then, as TFP returns towards steady state, households begin eating down the extra capital that they had accumulated. When households are less willing to intermporally substitute consumption, they substitute more on the leisure margin, as can be seen by the solid yellow line in the figure. Hours rise on impact, but quickly fall below steady state levels.

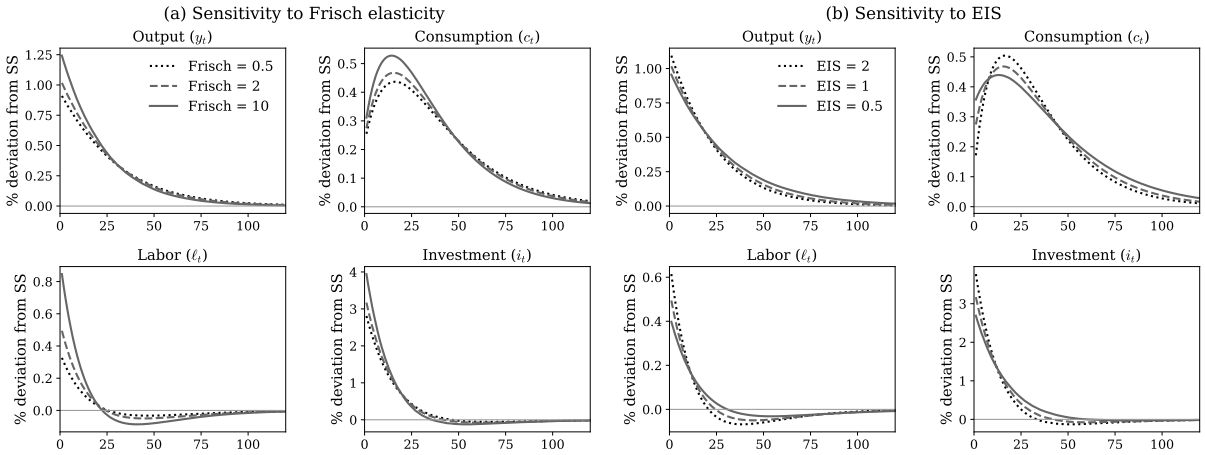


Figure 14.10: Sensitivity to EIS and Frisch Elasticity in RBC Model

To sum up, evaluated in the context in which the original RBC papers were published, the model was very successful at generating business cycle fluctuations and co-movements. The model explained nearly 80% of fluctuations in output, and replicated the relative volatilities of investment and consumption. That being said, the model still featured too much consumption smoothing and hours that were not volatile enough. Several extensions to the core framework (some of which are discussed later in this chapter) were developed to address some of the shortcomings of the model.

14.5 Extensions to the RBC Model

In this section, we briefly discuss some of the most prominent extensions to the RBC model, many of which remain as the building blocks for modern business cycle analysis.

14.5.1 Extensions of the RBC with indivisible labor

One of the earliest extensions of the RBC model introduced indivisible labor, addressing a key critique: the assumption that households can adjust work hours smoothly at business-cycle

frequencies. This frictionless adjustment was increasingly questioned by applied microeconomists. As discussed earlier in this chapter the implied Frisch elasticity by the calibration was approximately 2. Early empirical estimates using micro data (e.g., MaCurdy, 1981 and Altonji, 1986a) suggested the Frisch elasticity was close to zero.[12] The "micro" Frisch elasticity appeared small, but at the macro level the Frisch elasticity needed to be much higher to match the data.

A second critique was that there was no notion of unemployment or inactivity: everyone was always fully employed at the hours that she wished to work. At the same time, the data suggest that much of the aggregate hours adjustment in the data actually occurs along the extensive margin—because of fluctuations in aggregate employment—not because of hours conditional on being employed. Hansen (1985) and Rogerson (1988) introduced indivisibilities into the labor supply choice. They made the simple assumption that a consumer could either work full time, supplying $\bar{\ell}$ hours, or not work at all. The discrete choice introduces potential non-convexities, which would complicated the dynamic programming problem of the household. To simplify the analysis, they assumed that people had access to lotteries to convexify their labor supply choice. As discussed in Chapter 12, the model essentially collapses to one with an infinite aggregate Frisch elasticity at the extensive margin, since the disutility of hours is linear. The model also features a low elasticity of labor supply at the intensive margin—zero by construction in this case—which helps reconcile large fluctuations in employment with relatively small changes in hours worked among those employed.

14.5.2 RBC model with capital adjustment costs

A second shortcoming of the core RBC model was that investment was too volatile relative to output. One way to dampen the investment response to TFP shocks is to introduce capital adjustment costs to explain gradual investment behavior (Hayashi, 1982). Costs represent the real resource costs incurred by firms when changing their capital stock and are meant to capture various factors, including installation and setup costs, costs related to training workers to use new technologies, and costs associated with selling old equipment.

The investment adjustment cost shows up in the capital accumulation equation:

$$k_{t+1} = (1-\delta)k_t + i_t - \Psi(i_t, k_t)$$

where $\Psi(i_t, k_t)$ is the functional form for the capital adjustment costs. Typically, we assume that adjustment costs are convex (to keep the optimization problem well behaved), and take a quadratic form[13]

$$\Psi(i_t, k_t) = \frac{\psi}{2}\left(\frac{i_t}{k_t} - \delta\right)^2 k_t.$$

The idea is that maintain the capital stock at its previous level by replacing the depreciated capital is costless (e.g., maintenance as opposed to buying new machines). The quadratic

[12] One limitations of these early studies was that they focused on prime-age males. Women and younger and older men tend to have more elastic labor supply.

[13] Note also that the formulation is homogeneous of degree 1 in (i, k), so as to be consistent with zero profits for firms.

formulation yields a shadow price of "installed" capital equal to

$$1 + \psi\left(\frac{i_t}{k_t} - \delta\right),$$

which we define as q_t: Tobin's q. Firms increase investment when the price of capital exceeds that of output (or consumption), and reduce it otherwise.

When firms face these costs in adjusting their capital stock, they will spread their investment decisions over time. This can lead to more persistent and amplified responses to shocks, helping the model better match the observed persistence and magnitude of business cycles. All New Keynesian models with capital also feature adjustment costs to dampen the response of investment to movements in the real rate.

14.5.3 RBC model variable capacity utilization

Variable Capacity Utilization was introduced by Greenwood et al. (1988) and further explored by others. The basic idea is that firms can choose how intensively to use their capital stock depending on current prices for their goods. You can imagine that machines could be run for multiple shifts of workers, up to 24 hours a day. Or they could be used for one eight hour shift and then sit idle the other sixteen hours of the day. Variable capacity utilization allows for more flexible responses to shocks, as firms can adjust capital usage without immediate investment. The effective capital in the production function is thus modified as $u_t k_t$, at the cost of higher depreciation for higher capacity utilization:

$$k_{t+1} = (1 - \delta(u_t))k_t + i_t.$$

The idea is that if you utilize your capital more intensively today, then it will depreciate faster. Variable capacity utilization allows another margin of adjustment in response to TFP shocks. Now, firms can increase their effective capital services in addition to hiring more labor. This leads to greater amplification of TFP shocks, as compared to the standard model. The measurement of TFP shocks, however, now needs to be adjusted, since our Solow residual approach assumed a constant depreciation rate in constructing the capital stock series.

14.5.4 RBC model with additional shocks

In the standard RBC model, there is only one shock, z_t. As a result, all variables are highly correlated with each other. One way to break the high correlation between all of the variables is to introduce additional shocks. Here we discuss two of the most common additional shocks in the RBC literature, investment-specific technology shocks and government spending shocks.

RBC Model with investment-specific technology shocks These shocks affect the efficiency of new capital goods. They separate the technology affecting consumption goods production from the technology affecting investment goods production, allowing for differential productivity growth rates. The concept was introduced in a series of papers by

Greenwood et al. (1997); Greenwood, Hercowitz, and Krusell (2000) who looked at both the long-run and business cycle implications of investment-specific change; the long-run part is discussed in Chapter 13. Proposing investment-specific productivity shocks was motivated by the negative correlation between the price and quantity of investment in equipment at long-run and business cycle frequencies (the opposite of what one would expect from the adjustment cost model discussed above). In addition, investment-specific shocks can break the strong correlation between consumption and investment in the core model. The capital accumulation equation changes as follows:

$$k_{t+1} = (1 - \delta)k_t + q_t i_t$$

, where q_t is the investment-specific shock. They found that investment-specific shocks can account for roughly 30% of business cycle fluctuations.

Government spending shocks Another common shock introduced is a government spending shock. The shock can help break the high correlation between wages and hours, in which is at odds with the data. Let g_t indicate per capita government expenditure, so we modify the resource constraint from the core model to read:

$$c_t + i_t + g_t = y_t.$$

Why does introducing government consumption help break the correlation between hours and wages? Suppose households care about an aggregate consumption $C_t = [\theta g_t^{\varphi} + (1 - \theta)c^{\varphi}]^{1/\varphi}$, where φ measures the elasticity of substitution between c and g. If the elasticity is very high, government spending is like private consumption, and it will not have any effect on the outcomes we've looked at so far. On the other hand, if this elasticity is low, government spending is more like a tax and it reduce available resources, increasing labor supply due to negative income effect.

14.6 Business cycle accounting

In the previous section, we explored several frictions and extensions to the RBC model aimed at improving its empirical performance. While each brought the model closer to the data, it remains useful to have a systematic framework for assessing which features of the model align with the data—and which do not. In this section, we introduce such a methodology. Chari, Kehoe, and McGrattan (2007) propose Business Cycle Accounting (BCA) as a methodology that can be used to assess how different economic frictions (such as technology shocks, labor market frictions, or fiscal policy variations) contribute to business cycle dynamics. We can think of the frictions as deviations from the behavior that would be implied by the core RBC model. Their motivation was to provide a structured way to evaluate different frameworks (including extensions to RBC) based on their ability to explain observed economic fluctuations. The approach introduces four broad potential frictions into the RBC model: efficiency wedges (technology), labor wedges (labor market), investment wedges (capital investment), and government spending wedges (fiscal policy). These wedges

can be interpreted as distortions from the planner's allocation (recall that the equilibrium in the core RBC model is efficient) that could arise from various frictions or policies.

The wedges amount to modifications of the equations in the dynamic programming problem faced by the households that show up as potential distortions:

1. **Efficiency Wedge (z_t)**: This wedge affects the production function, which shows up as a TFP shock:
$$Y_t = z_t F(k_t, \ell_t).$$

2. **Labor Wedge ($1 - \tau_t^\ell$)**: This wedge shows up in the household's budget constraint as a tax on labor earnings and distorts the labor-leisure choice and can be seen as distortions in the labor market. Equation (14.8) is modified as:
$$(1 - \tau_t^\ell) z_t F_2(k_t, \ell_t) = \frac{U_2(c_t, 1 - \ell_t)}{U_1(c_t, 1 - \ell_t)}.$$

3. **Investment Wedge ($1/(1 + \tau_t^i)$)**: This wedge, which multiplies investment in the budget constraint, affects capital accumulation equation and can be interpreted as frictions in the investment sector. Equation (14.11) is thus modified as:
$$U_1(c_t, 1-\ell_t) \left[1 + \tau_t^i\right] = \beta \mathbb{E}\left[\left(z_{t+1} F_1(k_{t+1}, \ell_{t+1}) + (1 - \delta)\left[1 + \tau_{t+1}^i\right]\right) U_1(c_{t+1}, 1 - \ell_{t+1}) | z_t\right].$$

4. **Government Consumption Wedge (g_t)**: This variable affects the resource constraint and can be seen as reducing output (and in this sense is a distortion), given that utility is not affected by it.
$$c_t + i_t + g_t = Y_t.$$

In addition, we have to modify the household budget constraint with lump-sum transfers T_t, so that the aggregate resource constraint holds in equilibrium.

In practical terms, the efficiency wedge can capture a variety of phenomena, including technological changes, market power, changes in regulatory environments, and other factors that affect how efficiently an economy can convert inputs into outputs. In the framework, it will show up identically to a change in TFP that we've considered thus far. The labor wedge captures discrepancies between the marginal rate of substitution between consumption and leisure and the marginal product of labor. Factors contributing to the labor wedge include search frictions, taxation (see Chapter 15), and other policies affecting labor supply and demand, such as unemployment insurance or minimum wage laws. A significant labor wedge indicates that the labor market is not operating efficiently, potentially leading to underemployment or an inefficient allocation of labor resources. The frictions underlying the labor wedge will be explored further in Chapter 20 that introduces frictional labor markets and search.

The investment wedge is a measure of any distortions affecting the relationship between the cost of investment and the return on investment. It can arise from financial frictions (see Chapter 19), taxation on capital income (see Chapter 15), and policies affecting savings and investment decisions. An adverse investment wedge means that investment in the economy

is lower than what would be expected based on the fundamentals of the economy, leading to underinvestment in productive capital and, consequently, slower economic growth.

Finally, the government spending wedge affects the allocation of resources between private and public sectors. It reflects the impact of government expenditure (and taxation policies to fund this expenditure) on the economy's productive efficiency and the private sector's consumption and investment decisions. A government spending wedge might arise from government spending either crowding out private investment by raising interest rates or by reallocating resources in ways that are not aligned with market efficiency.

In order to implement the BCA approach, we first calibrate the preference and technology parameters of the model. Here, Chari et al. (2007) follow a similar procedure to the one we discussed above for calibrating the core RBC model, using the same functional forms and choices for the intertemporal elasticity of substitution and Frisch elasticity. Next, they assume that the four aggregate shocks (the wedges) follow an AR(1) process. Then they log-linearize the model with respect to the aggregate state and estimate the stochastic process for the wedges via maximum likelihood comparing the model generating time series for output, hours, investment, and government consumption from both the data and the simulated time series generated by the model. Armed with the estimated time-series and stochastic process for the wedges, they can then evaluate the consequences of the various wedges in isolation (or combinations of the four).

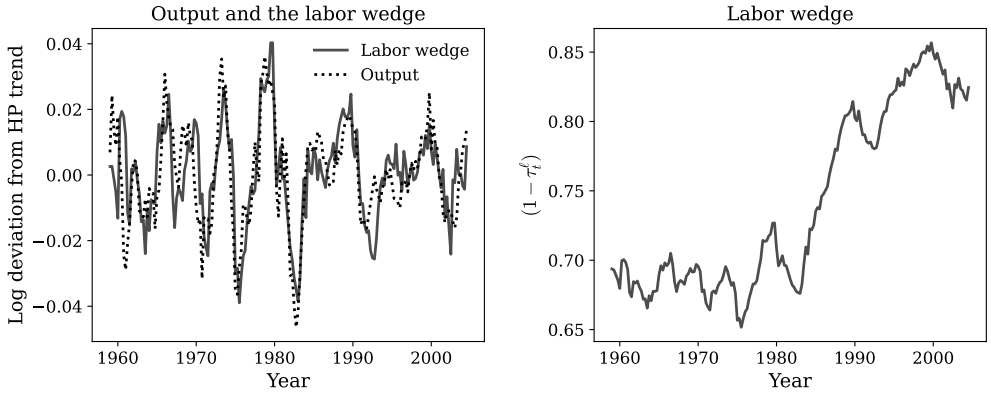


Figure 14.11: The estimated labor wedge from BCA

The outcome of the BCA estimation is that the efficiency wedge and the labor wedge are the two most important wedges for driving business cycle fluctuations: about 70% of the variance in output fluctuations are driven by the efficiency wedge and 25% by the labor wedge. These findings hold even when the authors look at the Great Depression (the main exercise was based on post-war data). The findings suggest that the potentially most promising extensions to the RBC model are ones that introduce explicit frictions in the labor market. The estimated labor wedge in log-deviations from an HP-filtered trend and in levels are plotted in Figure 14.11. Focusing first on the left panel, we can see that the labor wedge and output are highly correlated and have nearly the same standard deviation over the estimation sample: when output is low, the "tax" on working is high, making the wedge high: it is as if people like working less in recessions. Clearly, labor-market frictions and involuntary unemployment is a plausible interpretation of this finding, but not the only one.

Turning to the right panel, the level of the labor wedge has increased significantly over time (meaning less distortions in the labor market). The timing of the secular change in the labor wedge coincides with the changing nature of fluctuations that we saw in the data post-1985.

Recent work on "jobless recoveries"

The labor wedge itself explains nearly 75% of the variance in fluctuations in labor, suggesting that the labor wedge is potentially related to the phenomena of "jobless recoveries" that we discussed earlier in the Chapter. What could explain both the cyclical and secular trends in the labor wedge? Two recent theories have emerged that explain "jobless recoveries" based on real phenomena. One, proposed by Mitman and Rabinovich (2019), posits that countercyclical and unemployment-dependent extensions of unemployment benefits act as a time-varying policy wedge. Using a framework with frictional labor markets similar to that developed in Chapter 20, they show that benefit extensions explain roughly 1/3 of post-war labor market dynamics, and can explain the emergence of jobless recoveries (even though the underlying parameters of the model and shock processes remain constant).

Another explanation points to a striking trend in the U.S. around the same time period: namely the increase in female labor-force participation. The increase was particularly pronounced for married women. A series of recent papers (Olsson et al., 2019; Fukui, Nakamura, and Steinsson, 2023; Albanesi, 2019) thus show that for men, even the recessions before 1991 were "jobless", but masked in the aggregate by increasing female labor-force participation along the extension margin. The secular increase in participation can be explained by the decline in the gender-wage gap, which would mimic the secular trend in the estimated labor-wedge. When the secular increase in female participation plateaus in 1990, the jobless recoveries emerge in the aggregate. These facts can be rationalized in an extension to the RBC model with indivisible that takes gender and household composition seriously.

These examples help illustrate why the core of business cycle research has moved beyond the standard RBC model. Frictions in the labor market and household heterogeneity have emerged as important features in explaining business cycle phenomena (see, e.g., Krusell et al., 2017, Krueger, Mitman, and Perri, 2016). These will be discussed in Chapters 20 and 21, respectively.

14.6.1 Current frontiers of business cycle research

More than 40 years after the publication of Kydland-Prescott, business-cycle research continues to thrive. The COVID-19 pandemic, ensuing recession, and subsequent global surge in inflation have made clear that the period of tranquillity leading up to the Great Recession—known as the Great Moderation—is over. As predicted by Lucas (1980):

> "One would expect developments to arise from two quite different kinds of forces... Of these forces the most important... consists of purely technical developments that enlarge our abilities to construct analogue economies... The second source of technical developments is changes in the questions we want models to answer..."

Researchers have continued to extend the core RBC framework to meet the economic challenges of the 21st century. In the decade since the Great Recession, we have seen progress in incorporating household heterogeneity into workhorse macro models. One of the critiques of modern macro was its focus on the representative agent. While real business cycle theory was built on micro foundations, the frameworks were inconsistent with extensive empirical micro evidence on household behavior. The research frontier is now focused on making the core RBC theory a *micro-consistent* framework, i.e., consistent with empirical microeconomic evidence on household and firm behavior, expectations, and outcomes (for consumption, an idea pioneered by Deaton 1992b). It combines incomplete markets at the household level, and frictional product and labor markets.

Chapter 15

Government and public policies

Marina Azzimonti, Jonathan Heathcote, and Kjetil Storesletten

15.1 Introduction

The government has a large impact on economic outcomes through fiscal policy, monetary policy, and regulation policy. In this chapter, we focus on fiscal policy; in particular, on taxes, government spending, and debt. After presenting a summary of how governments tax, spend and borrow in practice, we turn to theory and discuss how fiscal policy choices impact the competitive equilibrium allocation, and how to frame the problem of optimizing over those policy choices. Monetary policy is discussed in Chapter 18. Our discussion centers on developed economies, with a particular focus on the United States. Chapter 24 introduces fiscal policy in emerging markets.

In Chapter 6, we discussed conditions under which the First Welfare Theorem holds. When markets are complete and competitive and there are no public goods or externalities –i.e., there are no "market failures"– competitive equilibrium allocations are Pareto optimal. Why then, do governments intervene in the economy? There are three main rationales.

The first is that there are public goods, such as national defense, that the market cannot provide because there is no way to restrict the enjoyment of public goods to those households who choose to pay for them. There are other goods and services, like education and healthcare, that are not pure public goods, but whose consumption confers large positive externalities. For example, if my neighbors are vaccinated, they are less likely to make me sick. Thus, absent government involvement, education and healthcare might be under-consumed.

A second reason governments intervene is that markets are not complete and competitive, in part because of private information frictions. For example, it might be difficult to buy private unemployment insurance, or annuities that insure against longevity risk. Thus, there may be a role for the government to provide public unemployment insurance, or to fund a public pension system. It is also possible that absent government intervention, the economy might occasionally get stuck in an inefficiently depressed equilibrium because of frictions in private markets. Thus, the government intervened during the Global Financial Crisis in 2008, bailing out a range of financial institutions to avoid a cascade of bankruptcies, and cutting taxes to try to boost consumer confidence.

The third rationale for government intervention is redistribution. Market economies tend to generate substantial income inequality, as discussed in Chapter 11 (and later in Chapter 21). This inequality may be Pareto efficient, but taxing the rich in order to fund transfers to the poor will generate a more equal allocation of resources, and one that a majority of households might prefer. In many economies, transfers account for most of total government spending. We discuss redistribution in Section 15.7.1.

The impact of the government on equilibrium allocations depends not just on how much the government wants to spend, but also on how the government pays for that spending. In practice, the taxes that households and firms are required to pay depend on the choices they make about how much to work and earn, how much to consume versus save, and how much to invest. Thus the tax system distorts all those choices, ultimately reducing output. We will explore the effects of distortionary taxation by adding proportional capital and labor income taxes to a standard neoclassical growth model. Note that higher public consumption or higher transfers necessitates higher tax rates, implying larger distortions to private sector choices and lower efficiency. There is significant cross-country variation in total government spending and in the extent of redistribution through the tax and transfer system. That suggests societies differ in how they view the trade-off between the benefits of a more equitable distribution of resources or higher public good provision, versus the efficiency costs of higher and more distortionary taxes (see Figure 15.1).

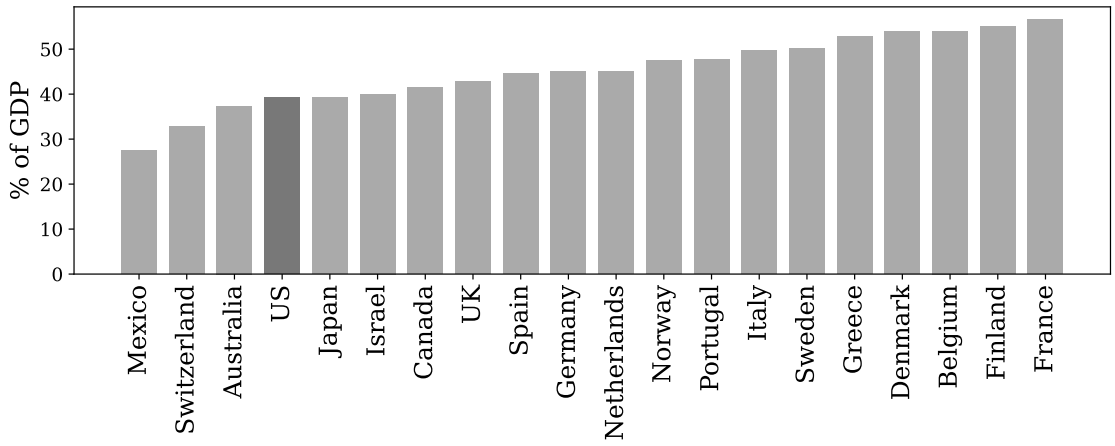


Figure 15.1: Government Spending across Countries (avg. 2010-2019).

15.2 Public Finance: An Overview of the Data

The government spends money on publicly-provided goods and services G_t, such as education and defense, and makes transfers T_t to individuals and corporations, such as food stamps and agricultural subsidies. These expenditures are financed out of tax revenues Rev_t, collected through taxes on goods and services (sales and excise taxes), taxes on income (income and payroll taxes), property taxes, and taxes on corporate profits. When revenues are insufficient to cover expenditures, the government borrows from domestic households and firms or from international lenders. Denoting the stock of debt at the start of period t by B_{t-1}, net

borrowing is equal to $B_t - B_{t-1}$. We denote the nominal interest rate on public debt by i_t, so $i_t B_{t-1}$ is interest payments. The government budget constraint can be written as

$$\underbrace{G_t + T_t + i_t B_{t-1}}_{\text{Expenditures}} = Rev_t + \underbrace{B_t - B_{t-1}}_{\text{Borrowing}}. \tag{15.1}$$

When expenditures – including interest payments – exceed revenues, we say that the government runs a *deficit*. In that case, $B_t > B_{t-1}$ and public debt rises. When expenditures are lower than revenues, the government runs a *surplus* and debt decreases. The stock of debt at any point in time, then, is the cumulative sum of net deficits run by a government through history. We now review some facts about the evolution of the main components of the government budget constraint to frame the topics discussed in this chapter.

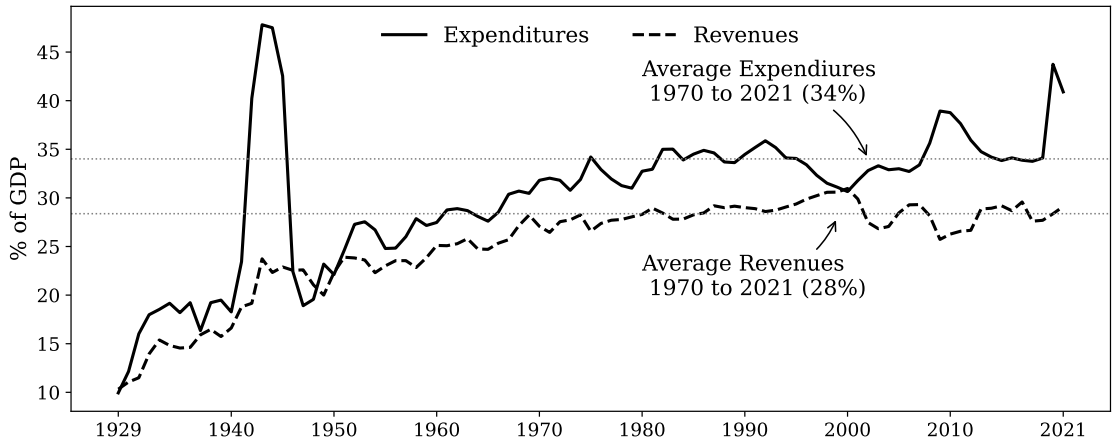


Figure 15.2: Revenues and Outlays, as percentages of GDP.

Figure 15.2 shows revenues and expenditures as percentages of GDP for the U.S. between 1929 and 2021. The series, which incorporate all levels of government (federal, state, and local), were obtained from the NIPA tables constructed by the Bureau of Economic Analysis (see Appendix 15.A.1 for details). Three key points can be drawn from this figure. First, both revenue and spending exhibit upward trends between 1929 and 1970, and then stabilize. Second, expenditures tend to exceed revenues, implying that the U.S. government typically runs deficits. Between 1970 and 2021 expenditures and revenues averaged 34 percent and 28 percent of GDP respectively. Third, expenditures jump during periods of war or recession, while revenues typically fall. Large increases in expenditure are evident during World War II, the Great Recession of 2007-2009, and the COVID-19 recession in 2020.

Figure 15.3 describes how the sources of tax revenue in the United States have changed over time. Over the post-war period, income and social insurance tax revenues increased significantly, and now account for around 11 and 7 percent of GDP, respectively. Revenue from sales and import taxes have been relatively constant at about 8 percent of GDP, while revenue from corporate taxes has declined and is now less than 2 percent of GDP.

The theoretical literature on the impact of taxes differentiates between taxes on labor income, on capital income, and on consumption.

In Section 15.3 we discuss the impact of these taxes on labor supply, investment, and

savings choices. Note, however, that this simple theoretical categorization does not map cleanly into the empirical partition of taxes: in particular, while labor earnings are part of the base for U.S. personal income taxes, income taxes taxes also apply to income accruing to capital, including unincorporated business income, dividend, interest, and rental income, and capital gains.

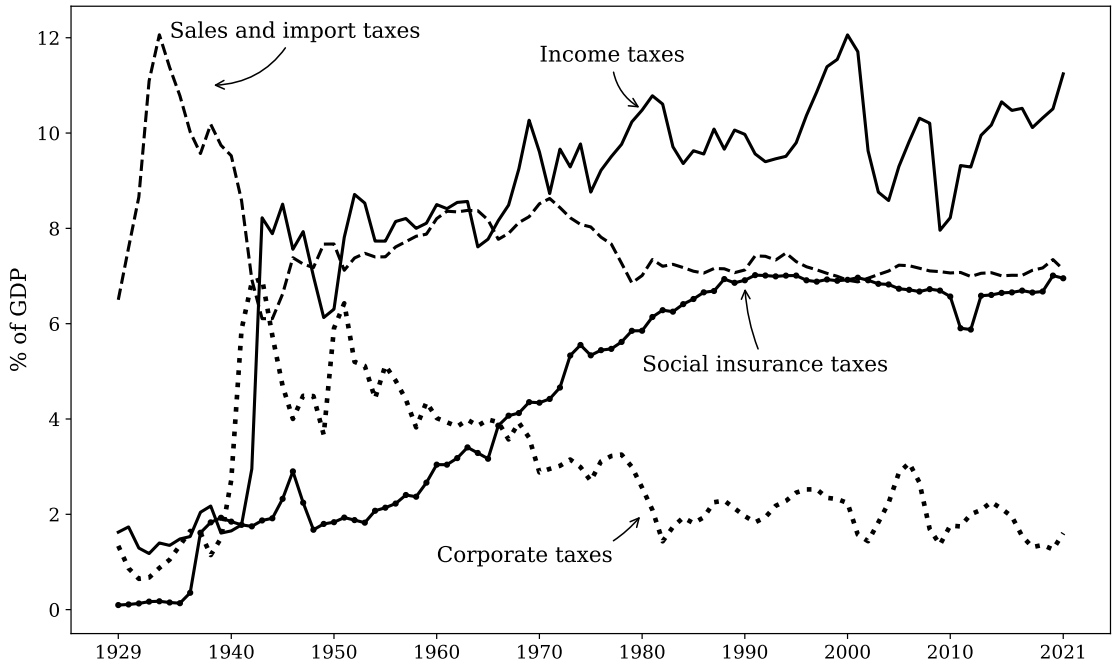


Figure 15.3: Taxes by Category, as percentages of GDP.

The government can postpone taxation by using public debt to finance expenditure. Does it matter whether the government finances spending out of current taxation versus whether it issues debt which must be repaid out of future tax revenue? In Section 15.4 we show that when taxes are lump-sum, the timing of taxes is irrelevant. This famous result is known as "Ricardian Equivalence." However, in the more realistic case in which taxes are distortionary, the timing of taxes does matter. What is then the optimal timing of taxes? In Section 15.5, we formalize the problem of a benevolent government that chooses a sequence for taxes and for debt to maximize social welfare, following a formulation known as the "Ramsey problem." Solving this problem we demonstrate an important "tax smoothing" result: it is optimal to finance temporary shocks such as wars, recessions or pandemics mostly by issuing debt. Evidence that governments do in fact smooth taxes over time is presented in Figure 15.4 (top panel), showing the total deficit (expenditures minus revenues, solid line), the primary deficit (which is the deficit excluding interest payments, dark bars), and interest payments (light bars). The U.S. government borrowed heavily during wars and recessions, particularly during WWII and the COVID-19 pandemic. The bottom panel shows the stock of debt. In the U.S., most public debt is issued by the Federal government, the result of balanced budget rules written into State constitutions. In other countries, a significant part of borrowing is done by sub-national units.

While borrowing is largest during recessions and wars, we also see that the U.S. has run

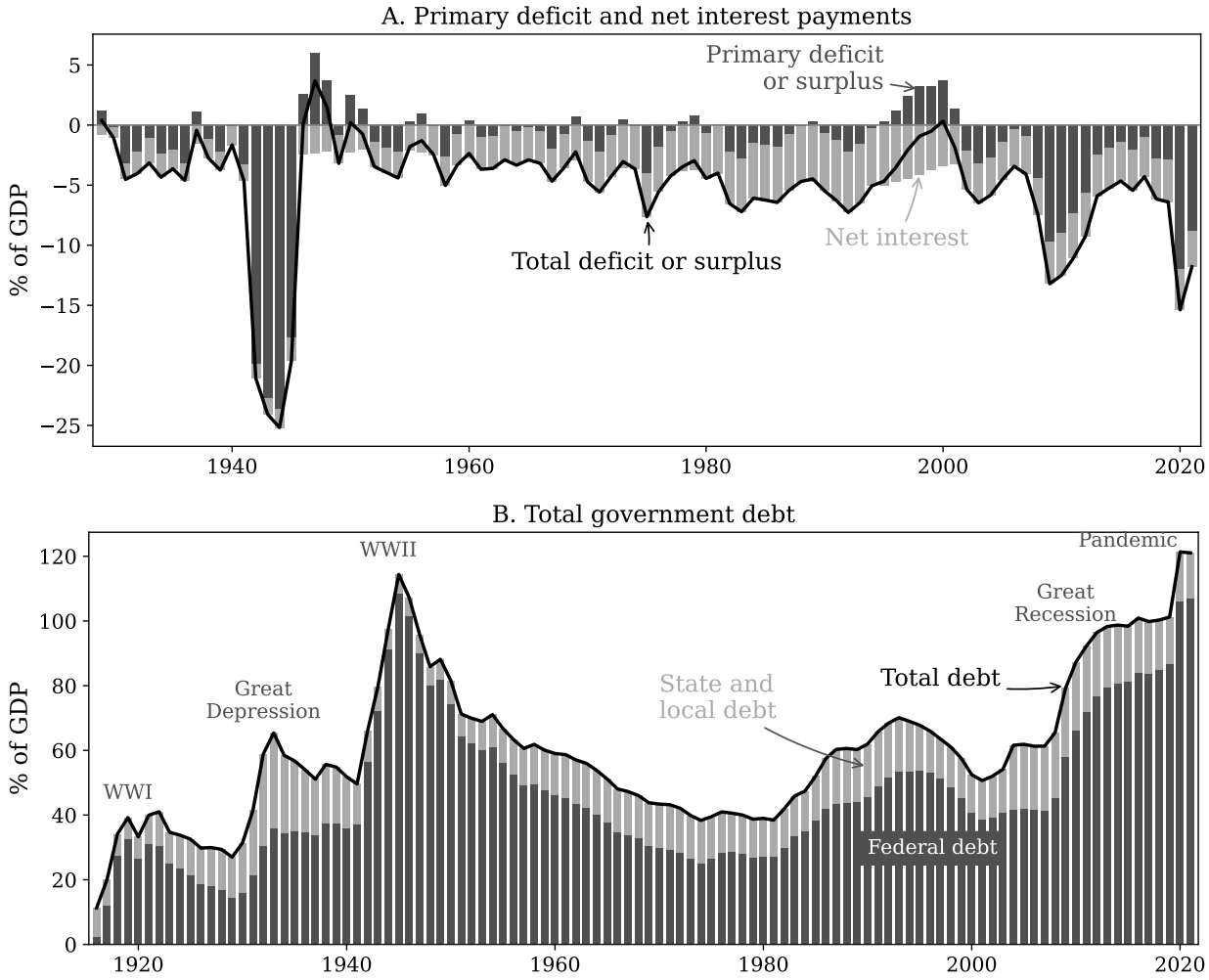


Figure 15.4: A: Total deficits, primary deficits, and net interest outlays. B: Total debt. All as percentages of GDP.

persistent deficits. This raises the question of how much debt is sustainable. One way to approach this question is to compare projections of future government deficits to the deficit levels that are consistent with a stable debt to GDP ratio.

Let D_t denote the primary deficit at date t (government spending excluding interest payments minus revenue). The government budget constraint (eq. 15.1) can then be written, in nominal terms, as

$$B_t = B_{t-1}(1 + i_t) + D_t.$$

Dividing through by nominal GDP at t gives

$$b_t = b_{t-1}\frac{1 + i_t}{(1 + \gamma_t)(1 + \pi_t)} + d_t,$$

where lower case letters denote values relative to nominal GDP, and where γ_t and π_t denote the growth rates of real GDP and the price level between $t - 1$ and t. Let $1 + r_t = (1 + i_t)/(1 + \pi_t)$ denote the *ex post* gross real interest rate between $t - 1$ and t. Thus, the debt

to GDP ratio evolves according to

$$b_t = b_{t-1}\frac{1+r_t}{1+\gamma_t} + d_t.$$

It is clear from this equation that the value of the real interest rate relative to the real growth rate is critical for the dynamics of public finances. When the primary deficit d_t is zero, the debt to GDP ratio will rise when $r_t > \gamma_t$, and will fall when $r_t < \gamma_t$.

One can ask what size primary deficit is consistent with a constant debt to GDP ratio. Debt will rise over time ($b_t > b_{t-1}$) if and only if

$$d_t > \frac{\gamma_t - r_t}{1+\gamma_t} b_{t-1}. \tag{15.2}$$

How much debt is sustainable in the U.S.?

At the time of writing (September, 2023) U.S. government debt held by the public is around 100 percent of annual U.S. GDP – i.e., $b_{t-1} = 1.0$. The growth rate of real GDP in the United States varies over time, but has averaged around 3 percent per year in the post-War period, suggesting $\gamma_t = 0.03$. The interest rate on 10 year inflation-protected government bonds is currently a little over 2 percent, suggesting $r_t = 0.02$.

Plugging these numbers into our debt sustainability equation (equation 15.2) suggests that the largest primary deficit consistent with debt not rising is approximately 1 percent of GDP. How does this compare to the actual primary deficit? The primary federal deficit in fiscal year 2022 was 3.6 percent of GDP, and the Congressional Budget Office (CBO) is forecasting primary deficits over the next 10 years of around 3.0 percent of GDP.[a] Thus, the U.S. debt to GDP ratio is likely to continue to grow. But will debt explode? Perhaps surprisingly, the arithmetic suggests not. In particular, given constant values for r and $\gamma > r$, any size primary deficit is consistent with a stable debt to GDP ratio, as long as that ratio is large enough. For example, suppose r_t and γ_t are expected to remain constant at values of 2 and 3 percent respectively. A 3 percent of GDP primary deficit is then consistent with a stable debt to GDP ratio of 309 percent of GDP (in terms of equation (15.2), $0.03 = \frac{0.01}{1+0.03} \times 3.09$). However, we should be very cautious about this calculation. As debt rises, the equilibrium real interest rate is likely to rise – investors will demand higher returns to buy all that debt. And once the differential between γ and r changes sign, stabilizing the debt to GDP ratio will require primary surpluses rather than deficits.[b]

[a] See Table 1.1 here: https://www.cbo.gov/publication/58946.

[b] See Hall and Sargent (2020) and Blanchard (2023) for more on debt sustainability. When debt burdens become large, countries sometimes choose to default (as discussed in Chapter 24).

The growth in the size of the U.S. government coincides with the creation (and expansion) of the Social Security system and the unemployment insurance program following the Great Depression, as well as the increase in public investment after WWII. This is illustrated by the top panel of Figure 15.5, which decomposes expenditures into the three sub-components

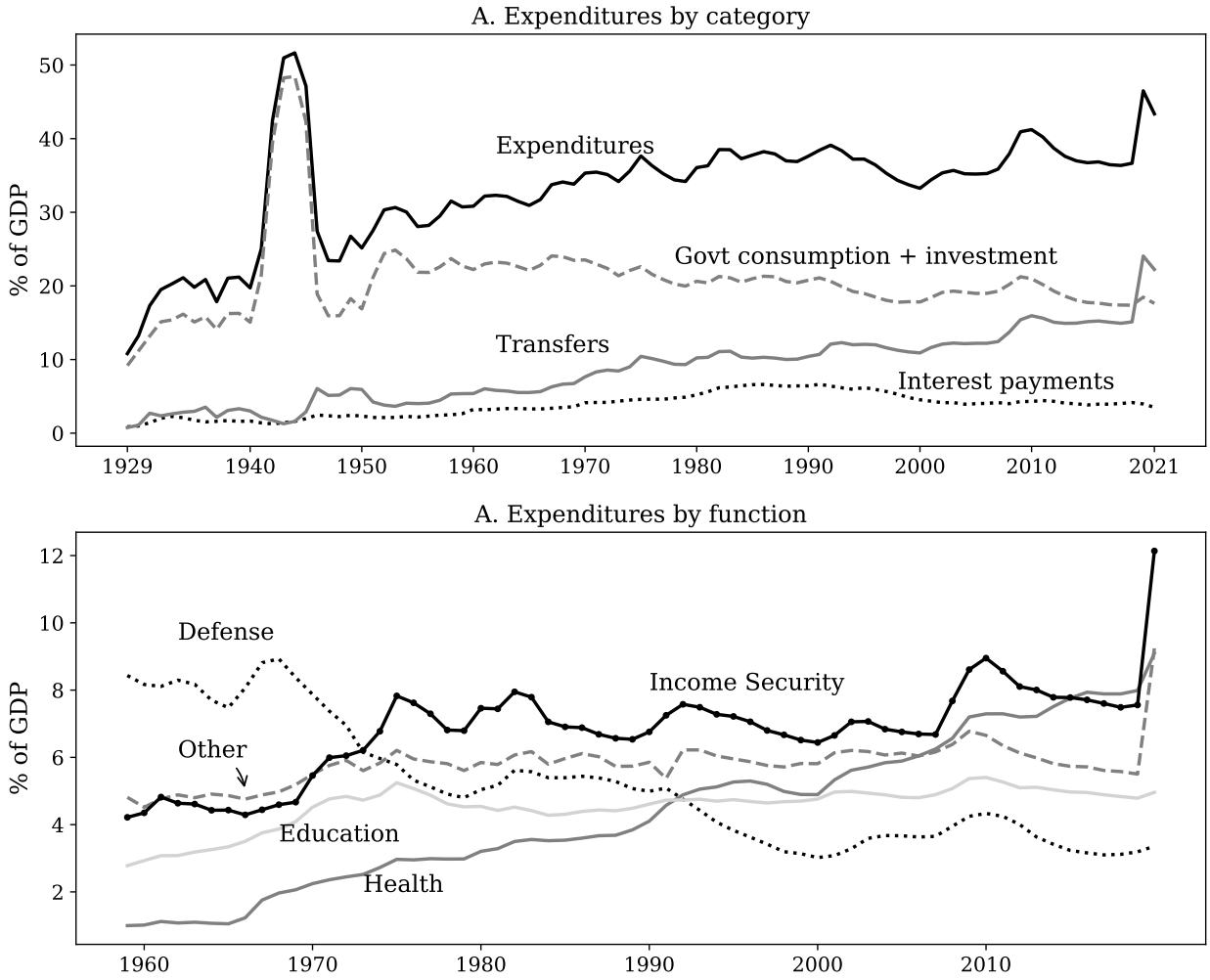


Figure 15.5: Expenditures by Category and function, as percentages of GDP.

shown on the left-hand side of the government budget constraint in equation (15.1). Early on in the sample, government consumption and investment constituted the largest portion of expenditures and drove most of the trend. Over time, transfers expand significantly, overtaking G_t during the COVID-19 pandemic. In the data, transfers include both redistributive and insurance programs. Redistributive policies will be studied in Section 15.7.

The bottom panel of the figure shows the evolution of expenditures by function for selected items from 1959 onward. Public education spending is relatively constant, accounting for 5 percent of GDP. Defense peaks in 1959 at 8 percent, but decreases significantly thereafter, to around 3 percent today. Health-care expenditures (including Medicare and Medicaid programs), on the other hand, show a steady rise, and now exceed 8 percent of GDP. Finally, "income security" spending, which includes unemployment insurance, retirement programs, disability and welfare, fluctuates around 7 percent of GDP. Spending on these items rises in recessions and decreases in booms, and as such these programs are typically referred to as "automatic stabilizers."

15.3 The effects of distortionary taxes

The objective of this section is to show how proportional taxes affect equilibrium allocations and prices. We do this in the context of the neoclassical growth model. Capital depreciates at rate δ. Households are infinitely-lived, discount at rate β, and enjoy utility each period from consumption and hours worked given by $u(c_t, \ell_t)$. They save in the form of capital, and rent capital and labor services to competitive firms at rates w_t and r_t. Firms produce according to a constant returns to scale production function $y_t = f(k_t, \ell_t)$. The government finances government consumption G_t and transfers T_t (which may be positive or negative) using proportional taxes on consumption, on labor income, and on rental income net of depreciation, $\tau_{c,t}$, $\tau_{\ell,t}$, and $\tau_{k,t}$. For now we assume no government debt (we introduce it in Section 15.4). The resource constraint is

$$C_t + G_t + K_{t+1} = f(K_t, L_t) + (1-\delta)K_t. \tag{15.3}$$

The government budget constraint is

$$G_t + T_t = \tau_{c,t}C_t + \tau_{\ell,t}w_tL_t + \tau_{k,t}(r_t - \delta)K_t. \tag{15.4}$$

The budget constraint for a representative household is

$$(1+\tau_{c,t})c_t + k_{t+1} = (1-\tau_{\ell,t})w_t\ell_t + k_t + (1-\tau_{k,t})(r_t-\delta)k_t + T_t. \tag{15.5}$$

A government policy is a sequence $\{\tau_{c,t}, \tau_{k,t}, \tau_{\ell,t}, G_t, T_t\}_{t=0}^{\infty}$.

Definition 15.1 *: A competitive equilibrium given a policy $\{\tau_{c,t}, \tau_{k,t}, \tau_{\ell,t}, G_t, T_t\}_{t=0}^{\infty}$ is a sequence of allocations $\{C_t, L_t, K_{t+1}\}_{t=0}^{\infty}$ and prices $\{w_t, r_t\}_{t=0}^{\infty}$ such that*

i. Given policy and prices, the sequence $\{C_t, L_t, K_{t+1}\}_{t=0}^{\infty}$ maximizes household lifetime utility $\sum_{t=0}^{\infty}\beta^t u(c_t, \ell_t)$ subject to budget constraints of the form (15.5) for all t, initial capital $k_0 = K_0$, and a borrowing constraint $k_{t+1} \geq 0$ $\forall t$.

ii. The allocation $\{L_t, K_t\}_{t=0}^{\infty}$ is a solution to the firm profit maximization problem at each date t, with $\ell_t = L_t$ and $k_t = K_t$ in equilibrium

$$\max_{k_t, \ell_t}\left\{f(k_t, \ell_t) - w_tk_t - r_tk_t\right\}.$$

iii. The government budget constraint eq. (15.4) is satisfied at each date t.

At each date t, the first order conditions that define optimal saving and labor supply decisions for a household are

$$\frac{1+\tau_{c,t+1}}{1+\tau_{c,t}}\frac{u_1(c_t, \ell_t)}{u_1(c_{t+1}, \ell_{t+1})} = \beta\left[1 + (1-\tau_{k,t+1})(r_{t+1}-\delta)\right] \tag{15.6}$$

and

$$-u_2(c_t, \ell_t) = \frac{1-\tau_{\ell,t}}{1+\tau_{c,t}} \cdot w_t \cdot u_1(c_t, \ell_t),$$

where where $u_i(\cdot,\cdot)$ is the partial derivative with respect to the ith argument, and $u_{ij}(\cdot,\cdot)$, $i,j=1,2$ is a second derivative.

The conditions for profit maximization are

$$w_t = f_2(k_t, \ell_t)$$

and

$$r_t = f_1(k_t, \ell_t),$$

where, similarly, $f_i(\cdot,\cdot)$ is the partial derivative with respect to the ith argument, and $f_{ij}(\cdot,\cdot)$, $i,j=1,2$ is a second derivative.

In order to discuss the distortionary effects of taxation, it is useful to compute the Pareto optimal allocation. Because this is a representative agent economy, the efficient allocation can be found by solving the problem of a benevolent planner that maximizes lifetime utility subject to the resource constraint

$$C_t + G_t + K_{t+1} = f(K_t, L_t) + (1-\delta)K_t.$$

The first-order conditions to this problem are

$$\frac{u_1(C_t, L_t)}{u_1(C_{t+1}, L_{t+1})} = \beta\left[1 + f_1(K_{t+1}, L_{t+1}) - \delta\right]$$

and

$$-u_2(C_t, L_t) = f_2(K_t, L_t)\cdot u_1(C_t, L_t). \tag{15.7}$$

Comparing across these two sets of conditions, we can see how taxes change households incentives to save and to work. Absent taxes, households equate the inter-temporal marginal rate of substitution between consumption at t and at $t+1$ to one plus the marginal product of capital, net of depreciation. With taxes, households care instead about the gross after-tax return to saving, which is given by

$$\frac{1+\tau_{c,t}}{1+\tau_{c,t+1}}\left[1 + (1-\tau_{k,t+1})(r_{t+1}-\delta)\right].$$

Holding r_{t+1} constant for a moment (in equilibrium r_{t+1} will depend on taxes) it is clear that taxes on rental income depress the after-tax return to saving, and that rising consumption taxes ($\tau_{c,t+1} > \tau_{c,t}$) work in the same direction. Similarly, taxes on labor income depress the return to working, as do taxes on consumption. Because taxes change workers' incentives to save and to work, they distort equilibrium allocations, and will typically reduce capital, labor supply, and output relative to the solution to the planner's problem.

15.3.1 Long-run distortions

Suppose we consider a steady state of the economy in which tax rates and allocations are constant. In such a steady state the household first-order conditions simplify to

$$1 = \beta\left[1 + (1-\tau_k)(r-\delta)\right] \tag{15.8}$$

and

$$-u_2(c,\ell) = \frac{1-\tau_\ell}{1+\tau_c} \cdot w \cdot u_1(c,\ell), \tag{15.9}$$

where

$$w = f_2(k,\ell) \tag{15.10}$$

and

$$r = f_1(k,\ell).$$

From the first of these it is immediate that a higher capital income tax τ_k must increase the steady state equilibrium rental rate for capital r. If the production function f has a Cobb-Douglas form, $f(k,\ell) = k^\alpha \ell^{1-\alpha}$, then $r = f_1(k,\ell) = (k/\ell)^{\alpha-1}$, and thus a higher rental rate corresponds to a lower capital-labor ratio. In particular,

$$\frac{k}{\ell} = \left(\frac{\alpha(1-\tau_k)}{\rho + \delta(1-\tau_k)}\right)^{\frac{1}{1-\alpha}}, \tag{15.11}$$

where $\rho = (1-\beta)/\beta$ denotes the household's rate of time preference. Note that because a higher τ_k depresses the steady state capital-labor ratio, it will depress the steady state wage w, in addition to raising r. In contrast, labor and consumption taxes have no impact on the pre-tax prices r and w.

The effect of taxes on labor supply will depend on the preference specification, via the marginal utility terms in eq. (15.9). An increase in τ_ℓ affects labor supply in the current period directly by reducing the after-tax wage, inducing ℓ to fall via a substitution effect. At the same time, because the individual becomes poorer when labor income declines, consumption shrinks, which results in a higher marginal utility of consumption, incentivizing the agent to work more via an income effect. The total effect of an increase of labor income taxes on ℓ is therefore ambiguous, depending on the relative strength of substitution and income effects.

For an illustrative example we consider a particular utility function, made famous by Greenwood, Hercowitz, and Huffman (henceforth, GHH):

$$u(c,\ell) = \log\left(c - \frac{\ell^{1+1/\phi}}{1+1/\phi}\right).$$

The GHH functional form is particularly tractable because consumption drops out of the first-order condition for hours worked; thus, this utility function can be described as one in which there are no income effects.[1] Steady state labor supply is given by

$$\ell = \left(\frac{1-\tau_\ell}{1+\tau_c}\right)^\phi w^\phi \tag{15.12}$$

Note that hours worked depend only on the return to working, where ϕ defines the elasticity of hours to after-tax wages. Higher labor income or consumption taxes depress hours worked. Higher capital income taxes also depress hours, via their negative impact on w.

[1] See Appendix 15.A.2 for an alternative utility function with income effects.

15.3.2 Tax incidence

In addition to simplifying the algebra, another feature of the GHH specification is that the steady state of the representative agent model specification is identical, at the aggregate level, to an alternative decentralization in which there are two household types: (i) workers, who rent labor services but own no capital, and (ii) capitalists, who own and rent out capital but who do not work. In what follows we focus on this worker-capitalist specification, because it allows for a discussion of tax incidence, namely the issue of who pays different sorts of taxes (in a representative agent setting, the representative household pays all taxes).[2]

From eqs. (15.10) and (15.12), we can solve for hours worked as a function of the capital to labor ratio

$$\ell = \left[\frac{1-\tau_\ell}{1+\tau_c}(1-\alpha)\left(\frac{k}{\ell}\right)^\alpha\right]^\phi$$

which, combined with eq. (15.11), gives an expression for steady state output

$$y = \left(\frac{k}{\ell}\right)^\alpha \ell = \left[\frac{1-\tau_\ell}{1+\tau_c}(1-\alpha)\right]^\phi \left[\frac{\alpha\left(1-\tau_k\right)}{\rho+\delta\left(1-\tau_k\right)}\right]^{\frac{\alpha(1+\phi)}{1-\alpha}}. \tag{15.13}$$

Note that all three tax rates affect the level of steady state output, and that the level of output is decreasing in each tax rate. Capital income taxes depress the capital labor ratio, but their impact on output is amplified by the fact that a lower capital-labor ratio means lower wages, which in turn depress labor supply.

Consider the case with no transfers ($T = 0$). In steady state, workers consume

$$c_w = \frac{1-\tau_\ell}{1+\tau_c} w\ell = \frac{1-\tau_\ell}{1+\tau_c}(1-\alpha)y,$$

while capitalists consume

$$c_k = \frac{1-\tau_k}{1+\tau_c}(r-\delta)k = \frac{\rho}{1+\tau_c}\left[\frac{\alpha(1-\tau_k)}{\rho+(1-\tau_k)\delta}\right]y.$$

From these expressions, it is clear that labor taxes directly depress the consumption of workers, while capital income taxes directly depress the consumption of capitalists. Consumption taxes depress the consumption of both types. Note, however, that all three types of taxes indirectly depress the consumption of both types via their impact on equilibrium output.

If we are designing a tax system, we would like to know more about how effective different sorts of taxes are in terms of raising revenue, relative to how distortionary they are in terms of depressing output. To make further progress on this question in a tractable way, we now make two additional assumptions. First, we temporarily rule out consumption taxes by setting $\tau_c = 0$. Second, we assume that the government has to devote a fraction g of

[2] Why is the steady state of the worker-capitalist model identical, given the GHH utility specification, to the steady state of the representative agent specification? The logic is that the level of consumption appears in neither the steady state first-order condition for saving, nor in the first order condition for hours worked. Thus the distribution of aggregate consumption between workers and capitalists has no impact on either steady state capital or steady state hours worked.

aggregate output to government purchases: $G = gY$. Thus, the steady state government budget constraint is

$$\begin{aligned} gy &= \tau_\ell w l + \tau_k (r - \delta) k \\ &= \tau_\ell (1 - \alpha) y + \tau_k \rho \frac{\alpha}{\rho + (1 - \tau_k)\delta} y. \end{aligned}$$

From this budget constraint we can immediately solve for the locus of budget-balancing pairs (τ_ℓ, τ_k) :

$$\tau_\ell = \frac{1}{1 - \alpha} \left[g - \tau_k \rho \frac{\alpha}{\rho + (1 - \tau_k)\delta} \right]. \tag{15.14}$$

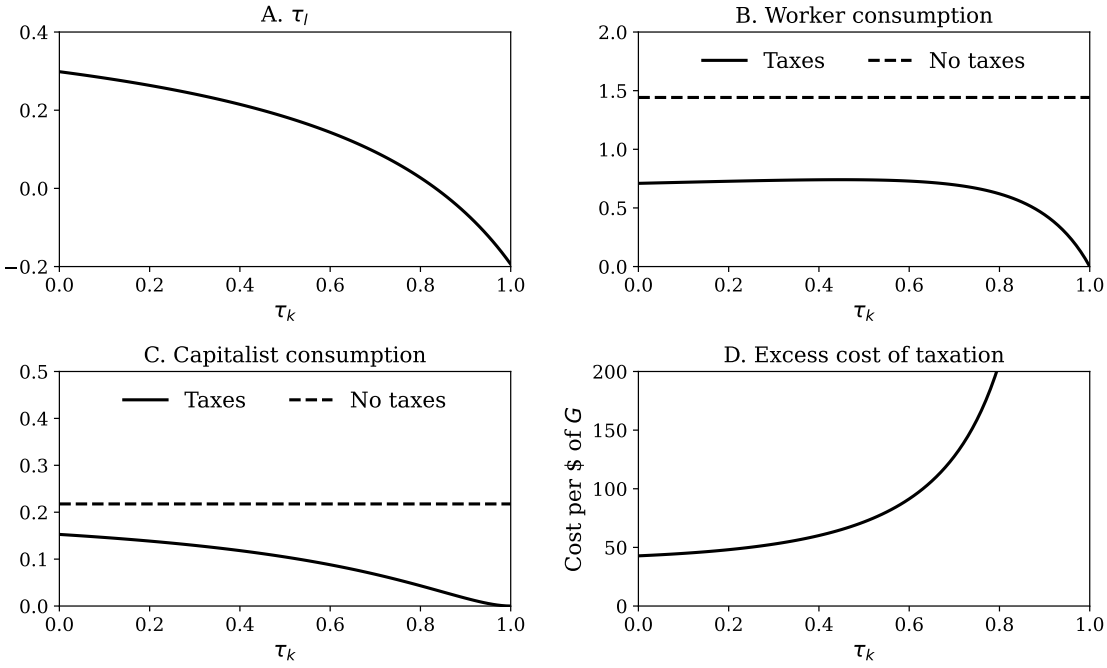


Figure 15.6: How Allocations Vary with τ_k.

Panel A of Figure 15.6 plots how the labor tax rate τ_ℓ varies with capital income tax rate τ_k according to eq. (15.14). The parameters used to construct the plot are $\alpha = 0.33$, $\delta = 0.07$, $\beta = 0.97$, $\phi = 1$ and $g = 0.2$. For each τ_k and the corresponding value for τ_ℓ given in Panel A, Panels B and C plot consumption of workers and capitalists, respectively. For comparison we also plot the levels of the two types' consumption for an economy without taxes. Note that as $\tau_k \to 1$, consumption of both types converges to zero. When τ_k is restricted to be non-negative, steady state consumption of capitalists is maximized at $\tau_k = 0$, while consumption of workers is a hump-shaped function of τ_k. Workers' utility depends on hours worked in addition to consumption, but given this utility function the consumption-equivalent argument of flow utility in equilibrium is proportional to consumption:

$$c_w - \frac{\ell^{1+1/\phi}}{1 + 1/\phi} = \frac{1}{1 + \phi}(1 - \alpha)(1 - \tau_\ell) y = \frac{1}{1 + \phi} c_w.$$

One way to define the deadweight cost of taxation is to ask by how much is total consumption reduced by taxes per unit of government consumption that the taxes finance. Let $c_{w,\tau=0}$ and $c_{k,\tau=0}$ denote the steady state consumption levels of workers and capitalists when $\tau_k = \tau_\ell = 0$, and define the excess cost of taxation as private consumption lost net of public consumption financed, measured per dollar of such spending:

$$\text{Excess Cost} = \frac{(c_{w,\tau=0} - c_w)/(1+\phi) + c_{k,\tau=0} - c_k - gy}{gy}.$$

If this ratio is equal to zero, then steady state utility in consumption units is reduced by one for each unit of government purchases. Panel D of Figure 15.6 plots the excess cost as τ_k varies (in the background τ_ℓ is adjusted with τ_k to balance the government budget constraint). When $\tau_k = 0.2$, each dollar of government consumption comes with an excess cost of 50 cents, meaning that total private consumption is reduced by \$1.50 for each dollar of public goods provided. Clearly, the excess cost of taxation is always positive. Furthermore, the excess cost of taxation is increasing and convex in the capital tax rate τ_k.

15.3.3 Tax reform

Panel D of Figure 15.6 suggests that the excess cost of taxation is minimized when $\tau_k = 0$. The tax rate on capital income in the U.S., however, ranges from 10% to 37% (the ordinary income tax brackets in 2022, applied to capital held less than a year). From Panels B and C of the same figure, we see that if capital taxes were reduced to zero and labor taxes were raised to support the same $g = G/GDP$ ratio, then capitalists' consumption in steady state would be much higher without much reduction in workers' consumption. These findings might suggest that eliminating capital income taxes would be a good idea. However, the steady state associated with a lower τ_k has a larger capital stock, and if capital taxes are reduced it will take time for the economy to accumulate this extra capital. Additional capital accumulation will come at the cost of reduced current consumption. It is therefore important to analyze *transitional dynamics* in addition to steady states when considering tax reforms. We illustrate this with a simple example, using the worker-capitalist framework from the previous section and again assuming no transfers, $T_t = 0$.[3]

We start from a situation in which $\tau_k = 0.25$, a midpoint of the current tax brackets. Labor taxes are set to $\tau_\ell = 0.2529$, obtained from eq. (15.14) to sustain $g = 0.2$ given the parameters used in Section 15.3.2. We assume that the economy is in steady state until period 10, and that a switch to $\tau_k = 0$ is implemented, unexpectedly and permanently, in period 11. At that date, we increase the labor tax to $\tau_\ell = 0.2985$ so that eq. (15.14) still holds at $g = 0.2$ in the new steady state. We allow G_t to vary during transition to balance the government budget date by date given constant tax rates. While the initial and final steady states can be characterized analytically, the evolution of k_t during transition requires the use of computational methods. We assume that the economy has reached the final steady

[3]While the steady state of the representative agent model (RA) is identical to the worker-capitalist environment (WK), transitional dynamics are not the same. In the RA model, the first order condition with respect to capital includes the disutility of labor (due to non-separability between c and ℓ), whereas this is absent in the WK environment, since capitalists set $\ell = 0$. In the example computed above, the difference is numerically insignificant. However, it could be sizable with other preference specifications.

state by date T. Knowing the initial and final conditions, k_0 and k_T, respectively, all we need is a sequence $\{k_t\}_{t=1}^{T-1}$ consistent with the first-order conditions of capitalists at dates $\{0, ..., T-2\}$. This is a system of $T-1$ inter-temporal first-order conditions and $T-1$ unknowns, which can be solved using a standard non-linear system of equations root-finding routine.[4]

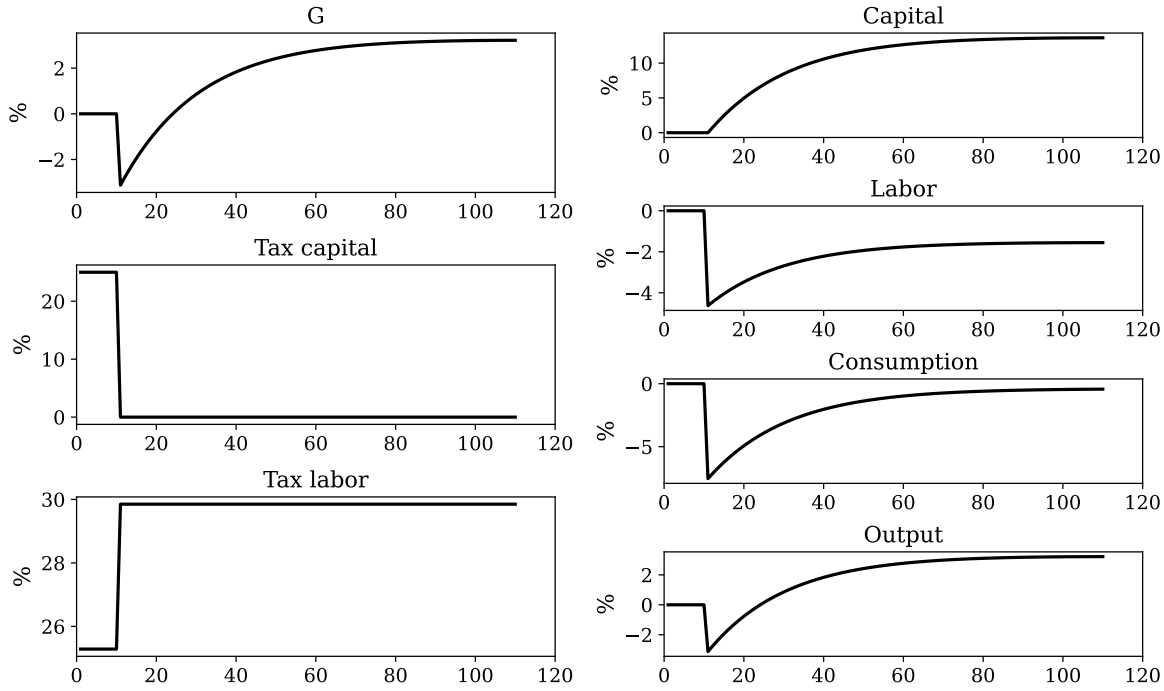


Figure 15.7: Eliminating capital income taxes.

Notes: Except for the tax rates, variables are plotted as percentage deviations from their values in the pre-reform steady state.

The left panel of Figure 24.4 displays the evolution of capital and labor income taxes (exogenous parameters to the model), as well as the endogenous evolution of G_t. Public spending changes in response to the reform because prices and allocations change in equilibrium (as seen in the right panel), in turn affecting government revenues (recall that G/Y is identical in the initial and final steady states). The elimination of capital income taxes encourages capital accumulation, whereas the increase in labor income taxes discourages labor supply upon impact.[5] Over time, capital grows, and because this positively affects wages, labor supply gradually recovers, ending up only slightly below the initial steady state level. GDP tracks labor supply in the short run, and capital in the long run, declining right after the reform and recovering slowly over time. Aggregate consumption decreases initially, both because output is low, and because lower capital taxes are stimulating saving and in-

[4]The values for consumption in the inter-temporal first-order condition can be substituted out using the budget constraint of capitalists, $(1+\tau_{c,t})c_k + k_{t+1} = k_t + (1-\tau_{k,t})(r_t - \delta)k_t$. It is important to verify that T is large enough that the economy has indeed converged to the new steady state.

[5]The latter is an artifact of the specific utility function used, since it exhibits no income effects. In Appendix 15.A.2, we re-compute this experiment using an alternative specification.

vestment. Subsequently, income rises and investment slows, pushing consumption back to a value close to the initial steady state.

While agents are eventually better off as the economy becomes more efficient (they work less than in the initial steady state but enjoy similar consumption), the exercise highlights that transitions can be painful. Whether the reform is beneficial for society overall depends on how much weight is assigned to capitalists versus workers. Capitalists are definitely better off, as their income always grows, whereas workers may be significantly worse off, as their tax burden increases (see, e.g., Domeij and Heathcote, 2004.) Even when considering a representative household, whether public spending is valued or not is important in this calculation. The reform is more desirable if agents derive utility from government consumption – which rises over time – than if public spending is entirely wasteful.

15.3.4 The Laffer Curve

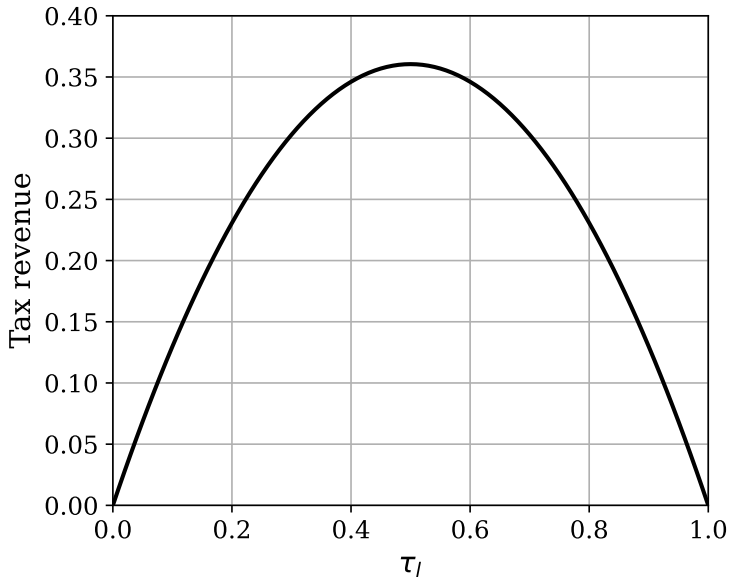


Figure 15.8: The Laffer curve

Another question of interest is: what combination of tax rates maximizes steady state tax revenue? A government that is fighting a war and wants to purchase the largest possible number of tanks might be especially interested in answering this question. When the planner only has access to taxes on labor earnings and rental income, and when tax rates must be positive, the revenue maximizing pair of tax rates is given by

$$\tau_\ell = \frac{1}{1+\phi}$$

and

$$\tau_k = 0.$$

Note that the higher is the Frisch elasticity of labor supply, ϕ, the lower is the revenue-

maximizing labor tax rate.[6]

Figure 15.8 plots tax revenue as a function of τ_ℓ, holding fixed τ_k at $\tau_k = 0$. This type of plot was popularized by Arthur Laffer in the 1970s, and is thus called a Laffer curve. Revenue is hump-shaped in the tax rate, and for tax rates above the revenue-maximizing rate, raising rates reduces revenue, because the tax base shrinks faster than the rate increases. Such a situation is known as "being on the wrong side of the Laffer curve." To see why the Laffer curve must be hump-shaped, it is enough to observe that at $\tau_\ell = 0$, no revenue is raised because no taxes are levied, while at $\tau_\ell = 1$ no revenue is raised because hours worked and output are equal to zero, and thus there is nothing to tax.

15.3.5 Theories of G

So far, we have assumed that public spending is exogenously given and generates no benefits to society; revenues are "thrown into the ocean." However, as shown in Figure 15.5, public expenditures are composed of government consumption, public investment, transfers, and interest payments. Here, we briefly describe theories of government consumption and public investment. Interest payments will be described in the next section, after we introduce debt into the model, and transfers and welfare programs will be described at the end of the chapter.

First, let us focus on how to model government consumption. We typically assume that the government has the technology to provide public goods that are valued by society. These include defense, law and order, taking care of parks and common areas, sanitation, etc. A key assumption is that the government uses resources to produce these goods, which then provide utility to agents. We typically assume that agents derive utility from public and private goods, $u(c, g)$, with $u_2(c, g) > 0$ and $u_{22}(c, g) \leq 0$, where $u_i(\cdot, \cdot)$ is the partial derivative with respect to the ith argument, and $u_{ij}(\cdot, \cdot)$, $i, j = 1, 2$ is a second derivative. The first best solution (e.g., when the government has access to lump-sum taxation) prescribes equating the marginal utility of private consumption to the marginal utility of public consumption: $u_1(c, g) = u_2(c, g)$. When taxes are distortionary, the government must take into account, in addition, the deadweight losses associated with taxation.

A second strand of theories of government spending focuses on the role of the government to provide key infrastructure such as roads, bridges, and schools. Letting k_g denote 'public capital,' the neoclassical growth model augmented to include public investment involves a production function,

$$f(k, k_g, \ell) = Ak^\alpha \ell^{1-\alpha} k_g^\theta,$$

[6] These rates are the solution to the problem:

$$\max_{\tau_\ell \geq 0, \tau_k \geq 0} \left\{ \tau_\ell (1-\alpha) y + \tau_k \rho \frac{\alpha}{\rho + (1-\tau_k)\delta} y \right\},$$

where output y is given by eq. (15.13). The first-order condition with respect to τ_ℓ gives the solution

$$\tau_\ell = \frac{1 - \tau_k \frac{\rho\phi}{1-\alpha} \frac{\alpha}{\rho + (1-\tau_k)\delta}}{1 + \phi}$$

Given this value for τ_ℓ, tax revenue is declining in τ_k at $\tau_k = 0$, indicating that $\tau_k = 0$ is the revenue-maximizing rate when tax rates must be positive.

and a law of motion for k_g,

$$k_{g,t+1} = i_{g,t} + (1 - \delta_g)k_{g,t},$$

with $i_{g,t}$ denoting public investment in period t, and δ_g the rate of depreciation of public capital. The parameter θ controls the elasticity of output with respect to public capital. If $\theta > 0$, there are increasing returns to scale. Estimates of θ vary, from as low as 0.05 (see Leeper, Walker, and Yang, 2010) to as high as 0.39 (Aschauer, 1989). If revenue can be raised in a non-distortionary way, then it is optimal to equate the marginal products of private and public capital, net of depreciation, $f_{k,t+1} - \delta = f_{g,t+1} - \delta_g$.

15.4 Government debt and Ricardian Equivalence

We now add government debt to the analysis. To simplify the presentation, we abstract from capital, as this reduces the number of state variables. Using a two-period example, we present conditions under which debt is irrelevant. This result is known as "Ricardian Equivalence." We then introduce distortionary taxation and explain how the government can use debt to smooth out tax distortions over time. The optimal sequence of taxes solves what is known as the "Ramsey Problem."

Consider a two-period representative household model. The government can issue debt in period $t = 0$ and can levy lump-sum taxes or hand out lump-sum transfers in periods $t = 0$ and $t = 1$. Suppose that the government contemplates a lump-sum transfer T_0 in the first period, financed by issuing government debt B_0, which will be paid back by levying a lump-sum tax τ_1 in the second period.

The representative household has utility defined over consumption and hours worked in periods 0 and 1 given by

$$u(c_0, \ell_0) + \beta u(c_1, \ell_1).$$

Households choose labor supply, consumption and saving in each period, taking as given exogenous wages, w_0 and w_1, and an exogenous return to saving, r_0. Abstracting from other taxes, the budget constraints in the two periods are

$$c_0 + b_0 = w_0\ell_0 + T_0$$

and

$$c_1 = w_1\ell_1 + (1 + r_1)\, b_0 - \tau_1,$$

where b_0 is the amount of debt the household buys in the first period, and r_1 is the interest paid on that debt. Dividing the second equation through by $1 + r_1$ and adding it to the first equation expresses the budget constraint in present value form:

$$c_0 + \frac{c_1}{1 + r_1} = w_0\ell_0 + \frac{w_1\ell_1}{1 + r_1} + T_0 - \frac{\tau_1}{1 + r_1}.$$

Note that the value for debt purchases, b_0, drops out of the present-value version of the budget constraint. In addition, note that given a promised interest rate of r_1, the second period tax τ_1 will have to satisfy $\tau_1 = (1 + r_1)\, B_0 = (1 + r_1)\, T_0$. Thus, the tax and transfer terms in the present value budget constraint must sum to zero, regardless of the size of the

initial transfer T_0. Since the tax and transfer scheme does not affect the lifetime budget constraint, the household's optimal allocation of consumption and hours must be identical to that in the case of zero taxes and transfers ($T_0 = \tau_1 = B_1 = 0$). The household must therefore respond to the initial transfer T_0 by increasing savings by exactly T_0. This extra savings will (i) exactly match the additional supply of government bonds issued, and (ii) provide exactly enough second period income to pay the expected lump-sum tax τ_1. This neutrality result is an example of *Ricardian Equivalence*. Note that this result hinges on the assumptions that taxes are lump-sum, that households face no credit constraints, and that the households who get the transfers are the same ones that must repay the debt. The result does not hinge on there being no capital. Using the same logic, it is possible to show that the result extends to an infinite-horizon economy (see also Barro, 1974 and Heathcote, 2005).

15.5 Ramsey Taxation

It is traditional in public finance to assume that the government cannot impose lump-sum taxes. Why? In a representative agent economy, there is no reason not to impose lump-sum taxes: such taxes are a distortion-free way to raise revenue. But in practice, actual households differ widely in terms of their income. Some households are so poor that they could not afford to pay a moderate lump-sum tax. Equally importantly, many people would find it unfair if the poor were expected to pay as much tax as the rich. Thus, the literature has focused on taxes that are *proportional* to income (see Chamley, 1986 and Judd, 1985 for early examples, and more recently Straub and Werning, 2020). Of course, one could consider making taxes a more complicated function of income – and we shall do so shortly – but proportional is simple, and simplicity can be viewed as a virtue.

But even if taxes are proportional to income, there is no need to tax different types of income at the same rate. In particular, earned income (income from labor) can be taxed at a different rate to unearned income (income generated by wealth). And if the government can save or borrow by issuing government debt, then it can also choose how tax rates should vary over time. We now consider a simple two period model and ask how a government that seeks to maximize the welfare of a representative agent should optimally set proportional taxes.

The timing assumptions are as follows. The government moves first, and announces tax rates for both periods, which we label periods 0 and period 1. To start, we assume that the government commits to these tax rates, and does not have the ability to deviate at $t = 1$ from the policies announced at $t = 0$. Later we will discuss how the analysis might change if the government does not have this commitment power.

A tax plan is *feasible* if there is a competitive equilibrium characterized by allocations and prices such that (i) those allocations are optimal choices for households and firms given prices and the tax rates described in the plan, (ii) the government budget constraint is satisfied, and (iii) markets clear.

A tax plan is *optimal* if it is feasible and the associated competitive equilibrium maximizes the welfare of the representative household. The optimal tax plan is called the *Ramsey plan*,

and the associated equilibrium the *Ramsey equilibrium*.[7] In general, lots of different fiscal plans will be feasible, but only one will be optimal. There are two different approaches in the literature to solving for the Ramsey plan.

The first and more intuitive approach is to work with the full set of equilibrium equations and variables, and to conceptualize the planner choosing tax rates to maximize household welfare, internalizing how changes in tax rates will affect all equilibrium variables. This is called the dual approach.

An alternative approach, called the primal approach, is sometimes easier to implement (see Atkeson, Chari, and Kehoe, 1999). Under the primal approach, we think of the planner as choosing equilibrium allocations for consumption and hours directly, subject to two sets of constraints. The first set of constraints ensure that allocations are technologically feasible. The second set of constraints ensure that there exists a set of tax rates such that the allocation is the competitive equilibrium given those taxes. These second constraints are called the *implementability constraints*. No equilibrium prices or tax rates appear in the primal problem. Once one has solved the primal problem, one can back out the tax rates that decentralize the solution in a final step.

15.5.1 The primal approach to optimal taxation: A simple example

The best way to understand how the primal approach works is to consider a simple example economy. Consider, in particular, the following two period model. There is a representative household with utility defined over consumption and hours worked in periods 0 and 1 given by

$$u(c_0, \ell_0) + \beta u(c_1, \ell_1).$$

The representative household supplies labor to a representative firm that produces output according to

$$y_0 = A_0 \ell_0$$

and

$$y_1 = A_1 \ell_1,$$

where A_t denotes potentially time-varying labor productivity. Because labor markets are assumed to be competitive, equilibrium wages equal productivities:

$$w_0 = A_0 \tag{15.15}$$

and

$$w_1 = A_1.$$

Households (and the government) can save or borrow using a storage technology that converts one unit of output at date 0 into $1 + r$ units of output at date 1, where r is an exogenous constant.[8] Let b_{t-1} denote household wealth at the start of period t. Assume that $b_{-1} = 0$.

[7] After Frank Ramsey, who wrote a handful of important papers in economics and more in the fields of mathematics and philosophy, before his death in 1930 at the age of 26.

[8] One interpretation might be that the economy is small and open, and r is the world interest rate.

The government must finance exogenous expenditures g_0 and g_1 in periods 0 and 1. It can raise tax revenue via proportional taxes on labor income at rates τ_0 and τ_1, and by taxing income from wealth in period 1 at rate $\tau_{b,1}$.[9] A *tax plan* is a vector $\{\tau_0, \tau_1, \tau_{b,1}\}$.

Let $b_{g,0}$ denote government savings in the storage technology at date 0. The government budget constraints for periods 0 and 1 are

$$g_0 + b_{g,0} = \tau_0 w_0 \ell_0$$

and

$$g_1 = \tau_1 w_1 \ell_1 + \tau_{b,1} r b_0 + (1+r) b_{g,0}.$$

Note that government savings at $t = 0$ delivers income at $t = 1$. If $b_{g,0} < 0$, the government is borrowing. These two constraints can be combined to give

$$g_0 + \frac{g_1}{1+r} = \tau_0 w_0 \ell_0 + \frac{\tau_1 w_1 \ell_1}{1+r} + \frac{\tau_{b,1} r b_0}{1+r}.$$

Given taxes and wages, the representative household solves

$$\max_{\{c_0, \ell_0, c_1, \ell_1, b_0\}} u(c_0, \ell_0) + \beta u(c_1, \ell_1)$$

subject to

$$c_0 = (1 - \tau_0) w_0 \ell_0 - b_0$$

and

$$c_1 = (1 - \tau_1) w_1 \ell_1 + (1 + r(1 - \tau_{b,1})) b_0,$$

where again the two budget constraints can be collapsed to give

$$c_0 + \frac{c_1}{1 + r(1 - \tau_{b,1})} = (1 - \tau_0) w_0 \ell_0 + \frac{(1 - \tau_1) w_1 \ell_1}{1 + r(1 - \tau_{b,1})}.$$

The first-order conditions that characterize the solution to the household's problem are

$$u_{c,0} w_0 (1 - \tau_0) = -u_{\ell,0}, \tag{15.16}$$

$$u_{c,1} w_1 (1 - \tau_1) = -u_{\ell,1},$$

and

$$u_{c,0} = \beta (1 + r(1 - \tau_{b,1})) u_{c,1},$$

where $u_{c,t}$ denotes the marginal utility of consumption in period t and $u_{\ell,t}$ is the marginal utility of labor (which is a negative number) in period t.

[9] The household has no wealth at date 0, so a tax on wealth or income from wealth at date 0 would not raise any revenue. If the household did have wealth at date 0, the government would like to tax that wealth, since that would effectively amount to a non-distortionary lump-sum tax. In the spirit of not allowing for lump-sum taxation, it is typically assumed that the Ramsey planner cannot tax initial wealth, or that there is an upper bound on the feasible initial tax rate.

Resource feasibility in this economy can be summarized by a single equation, which states that the present value of private plus public consumption is equal to the present value of output

$$c_0 + g_0 + \frac{c_1 + g_1}{1+r} = A_0\ell_0 + \frac{A_1\ell_1}{1+r}. \tag{15.17}$$

What about the implementability constraints? An allocation is a competitive equilibrium if it satisfies the three first-order conditions from the household problem, the two equilibrium expressions for wages, and the household and government budget constraints. We now show that these six equations can be collapsed into a single implementability condition. The idea is to take the household lifetime budget constraint, and to use the household first-order conditions to substitute out for $(1-\tau_0)w_0$, $(1-\tau_1)w_1$, and $1+r(1-\tau_{b,1})$. In particular, the first-order condition for saving implies that, in any competitive equilibrium, it must be the case that

$$1 + r(1-\tau_{b,1}) = \frac{u_{c,0}}{\beta u_{c,1}},$$

while those for labor supply imply

$$(1-\tau_t)w_t = -\frac{u_{\ell,t}}{u_{c,t}}.$$

After these substitutions, the household lifetime budget constraint can be written as

$$c_0 + \frac{c_1}{\frac{u_{c,0}}{\beta u_{c,1}}} = -\frac{u_{\ell,0}}{u_{c,0}}\ell_0 - \frac{u_{\ell,1}}{u_{c,1}}\ell_1 \frac{1}{\frac{u_{c,0}}{\beta u_{c,1}}}, \tag{15.18}$$

or, after multiplying through by $u_{c,0}$, as

$$u_{c,0}c_0 + \beta u_{c,1}c_1 = -u_{\ell,0}\ell_0 - \beta u_{\ell,1}\ell_1.$$

This is the implementability condition. It can be shown that any allocation that satisfies the resource constraint and the implementability condition can be implemented by some feasible tax plan (see, e.g., Atkeson et al., 1999).

It should be clear that the implementability condition embeds the household optimality conditions for labor supply and savings in addition to the household budget constraint. But what about the government budget constraint? We do not have to worry separately about that, because if the resource constraint and the household budget constraint are satisfied, the government budget constraint must also be satisfied, by Walras Law.

The *Ramsey problem* is to maximize lifetime utility for the representative agent, subject to the resource and implementability constraints. Writing this problem as a Lagrangian, with multipliers λ and μ on the resource and implementability constraints, we have

$$\max_{c_0,c_1,\ell_0,\ell_1} \Bigg\{ \begin{array}{l} u(c_0,\ell_0) + \beta u(c_1,\ell_1) \\ \quad + \lambda\left(A_0\ell_0 + \dfrac{A_1\ell_1}{1+r} - c_0 - g_0 - \dfrac{c_1+g_1}{1+r}\right) + \mu\left(u_{c,0}c_0 + \beta u_{c,1}c_1 + u_{\ell,0}\ell_0 + \beta u_{\ell,1}\ell_1\right) \end{array} \Bigg\}$$

The first-order conditions are

$$u_{c,0} - \lambda + \mu u_{c,0} + \mu u_{cc,0} c_0 + \mu u_{\ell c,0} \ell_0 = 0,$$

$$u_{\ell,0} + \lambda A_0 + \mu u_{\ell,0} + \mu u_{c\ell,0} c_0 + \mu u_{\ell\ell,0} \ell_0 = 0,$$

$$\beta u_{c,1} - \frac{\lambda}{1+r} + \beta\mu u_{c,1} + \beta\mu u_{cc,1} c_1 + \beta\mu u_{\ell c,1} \ell_1 = 0,$$

and

$$\beta u_{\ell,1} + \frac{1}{1+r}\lambda A_1 + \beta\mu u_{\ell,1} + \beta\mu u_{c\ell,1} c_1 + \beta\mu u_{\ell\ell,1} \ell_1 = 0,$$

where, for example, $u_{cc,0}$ denotes $\partial u_{c,0}/\partial c_0$.

These four first-order conditions alongside eqs. (15.17) and (15.18) constitute six equations that can be used to solve for the six unknowns $(c_0, c_1, \ell_0, \ell_1, \lambda, \mu)$. Thus, one can solve for the Ramsey allocation. Consider, in particular, a separable utility function of the form

$$u(c, \ell) = \frac{c^{1-\sigma}}{1-\sigma} - \frac{\ell^{1+1/\phi}}{1+1/\phi}.$$

In this case, the cross derivative terms drop out, and the second derivatives simplify to

$$u_{cc,t} c_t = -\sigma u_{c,t}, \;\; u_{\ell\ell,t} \ell_t = \frac{1}{\phi} u_{\ell,t}.$$

Thus, the first-order conditions can be written as

$$u_{c,0}(1 + \mu - \mu\sigma) = \lambda,$$

$$u_{\ell,0}\left(1 + \mu + \frac{\mu}{\phi}\right) = -\lambda A_0,$$

$$\beta u_{c,1}\left(1 + \mu - \mu\sigma\right) = \frac{1}{1+r}\lambda,$$

and

$$\beta u_{\ell,1}\left(1 + \mu + \frac{\mu}{\phi}\right) = -\frac{1}{1+r}\lambda A_1.$$

Comparing the first and the third, it is immediate that, at an optimum

$$\beta(1+r) u_{c,1} = u_{c,0}.$$

Comparing the first and the second, we see that

$$u_{c,0} A_0 \frac{1 + \mu - \mu\sigma}{1 + \mu + \mu/\phi} = -u_{\ell,0}.$$

Similarly, the third and the fourth give

$$u_{c,1} A_1 \frac{1 + \mu - \mu\sigma}{1 + \mu + \mu/\phi} = -u_{\ell,1}.$$

Comparing these expressions to the first order conditions for saving and for working in the original economy (eqs. 15.16), and noting that, in equilibrium, $w_0 = A_0$ and $w_1 = A_1$, it is clear that the only way both sets of first order conditions can be satisfied at the same allocation is if

$$\tau_{b,1} = 0$$

and

$$\tau_0 = \tau_1 = 1 - \frac{1 + \mu - \mu\sigma}{1 + \mu + \mu/\phi} = \frac{\mu\,(\sigma + 1/\phi)}{1 + \mu + \mu/\phi}.$$

Thus, this simple example illustrates two classic results in the Ramsey taxation literature. First, the government should commit to neither tax nor subsidize income from savings (see Chamley, 1986 and Judd, 1985). Second, the labor tax rate should be constant over time, and will be positive as long as either g_0 or g_1 is strictly positive (so that revenue must be raised). This result is described as *tax smoothing*, and the idea is that because distortions from taxes increase with the tax rate in a convex fashion, constant tax rates are preferable to time-varying tax rates (see Barro, 1979 and Lucas and Stokey, 1983).

15.5.2 Time consistency

Let us assume that parameters are such that the household saves in period 0 under the Ramsey plan. One parameter configuration that would deliver this is $\beta(1+r) = 1$ – so that the Ramsey allocation features $c_1 = c_1$ – and $A_0 > A_1$, so the Ramsey allocation features $A_0\ell_0 > A_1\ell_1$.

Recall that our analysis above presumed that the government announced a tax plan at date 0 and stuck to the plan at date 1. Suppose now that we give the planner the ability to redesign taxes in period 1. At $t = 1$ the planner can raise revenue either by taxing labor earnings (as promised in the original plan) or by taxing household income from savings. What combination of taxes would a benevolent planner choose? The answer is that such a planner would set $\tau_1 = 0$ and $\tau_{b,1}$ as high as necessary to fund required government purchases. The reason is that once period 1 rolls around, household wealth b_1 is already determined, and taxes on income from wealth are effectively a lump-sum tax. In contrast, taxes on labor earnings are distortionary. Because the planner would like to deviate from the Ramsey plan at date 1, given the chance to do so, the plan is said to be *time inconsistent* (see Kydland and Prescott, 1977).

There are many policy questions where time consistency arises as a central issue. For example, Chapter 24 focuses on debt policy. If governments can commit to repaying debt, they will be able to borrow cheaply. But those promises to repay may not be time consistent, in the sense that once debt has been accrued the government might be better off defaulting.

Does the fact that the Ramsey plan described above is time inconsistent mean that we should not take it too seriously as a practical policy prescription? It is certainly useful to know what policy would be optimal given commitment and to understand how the planner might be tempted to deviate from the Ramsey plan. It might be possible to design institutions in such a way that the planner has more commitment power – for example, by writing a constitution that precludes frequent tax changes. At the same time, there is a large literature that attempts to characterize the best time consistent policies (see Klein, Krusell, and Ríos-Rull, 2008).

15.6 Debt and pensions with overlapping generations

We next study fiscal policy in a non-dynastic economy and return to the two-period overlapping-generations endowment economy discussed in Section 5.5. We extend this model to incorporate government debt, a pay-as-you-go (PAYG) pension system, and taxes. To simplify the exposition, we assume a small open economy with access to a global bond market.

The population grows at a constant rate n. Let $N_t = (1+n)^t$ denote the size of the newborn young population at date t. The share of young people is $N_t/(N_t + N_{t-1}) = (1+n)/(2+n)$. Only the young work and their endowment of efficiency units grows at rate γ. Thus, the labor income of an individual born in period t is $y_t = (1+\gamma)^t \omega$ and aggregate labor income is $Y_t = y_t N_t = (1+n)^t (1+\gamma)^t \omega$.

The government operates a PAYG pension system and provides a public good G_t. The pension system pays p_t to every old individual, financed by taxing labor income at rate τ_p. The pension per retiree is therefore

$$p_t = (1+n)\,\tau_p y_t.$$

Spending on the public good is assumed to be a fixed fraction of GDP: $G_t = gY_t$. Government spending is financed by taxing labor income by a flat tax τ_t and by issuing debt. We abstract from taxes on capital income. The government budget constraint is given by

$$G_t + p_t N_{t-1} + (1+r)\,B_{t-1} = (\tau_p + \tau_t)\,Y_t + B_t,$$

where B_t is the issuance of new debt that matures next period and the interest rate r is exogenous and fixed. Because the pension system is self-financing, the budget constraint can be expressed as follows,

$$g + \frac{1+r}{(1+n)(1+\gamma)} b_{t-1} = \tau_t + b_t,$$

where b_t is the debt to GDP ratio at the end of period t.

Individuals maximize discounted utility, $u(c_{y,t}) + \beta u(c_{o,t+1})$, subject to budget constraints when young and old:

$$c_{y,t} + a_t = (1 - \tau_t - \tau_p)\,y_t \tag{15.19}$$

and

$$c_{o,t+1} = (1+r)\,a_t + p_{t+1}, \tag{15.20}$$

where a_t denotes saving at t, and where, in equilibrium, $p_{t+1} = (1+n)(1+\gamma)\tau_p y_t$. The sequence for optimal consumption can then be computed from two equations: a lifetime budget constraint and an Euler equation:

$$c_{y,t} + \frac{c_{o,t+1}}{1+r} = \left[1 - \tau_t - \tau_p + \frac{(1+n)(1+\gamma)}{1+r}\tau_p\right] y_t \tag{15.21}$$

and

$$1 = (1+r)\,\beta \frac{u'(c_{o,t+1})}{u'(c_{y,t})}. \tag{15.22}$$

Note, first, that a pension system is equivalent to issuing a particular form of government debt. To see this, consider an individual who has no pension tax or transfer ($\tau_p = p_{t+1} = 0$) but who is forced to purchase $b_{p,t}$ government bonds with a promised return r_p. The budget constraints for this individual would be

$$c_{y,t} + a_t + b_{p,t} = (1 - \tau_t)\, y_t, \tag{15.23}$$

and

$$c_{o,t+1} = (1 + r)\, a_t + (1 + r_p)\, b_{p,t}. \tag{15.24}$$

If we set $b_{p,t} = \tau_p y_t$ and $1+r_p = (1 + n)\,(1 + \gamma) \approx (1 + n + \gamma)$, then the budget constraints with "pension debt" (15.23-15.24) are equivalent to those with a pension system, (15.19-15.20). Note that the return on forced saving in the PAYG pension system is equal to the growth rate of output. Therefore, the pension system increases the present value of household income for the young if and only if wage growth exceeds the interest rate, i.e., iff $\gamma + n > r$ (this condition determines whether the right-hand side of eq. (15.21) is increasing in τ_p). This insight has an important implication: if the interest rate is larger than the growth rate of output $n + \gamma$ (a "normal" scenario) then the pension system is effectively a tax on the young generation. The generation who are old when the system is first introduced gain because they receive benefits without having paid taxes when they were young. But the current young generation and all future generations lose because they get a higher return on private savings than on pension contributions. However, if the interest rate is *lower* than the wage growth rate $n + \gamma$, then all generations gain from introducing a pension – both the initial old generation and all generations of young. In this case, introducing a PAYG pension system is Pareto improving. This corresponds to an equilibrium featuring dynamic inefficiency, as discussed in Section 6.4.1.

A major change in this overlapping-generations model relative to the standard infinite-horizon model is that Ricardian equivalence no longer holds. In particular, the timing of taxes and transfers now matters for the distribution of consumption across cohorts, and for the trajectory of aggregate consumption. To see this, consider a one-time transitory tax holiday in period t in an economy with zero initial debt. Thus, $\tau_t = 0$ and $B_t = gY_t$. Moreover, assume that the government finances the repayment of this debt by increasing taxes in period $t + 1$ and does not issue debt thereafter. Note, first, that this "tax holiday" increases the present value of consumption for generation t and lowers the present value of consumption for generation $t + 1$; see equation (15.21). Thus, the debt-financed tax cut shifts the tax burden from generation t to generation $t + 1$. The result is that aggregate consumption will increase in period t and fall in period $t + 2$.[10] Thus, Ricardian equivalence breaks down. This is different from the case we studied in Section 15.4. There, Ricardian equivalence held because the government's debt policy did not affect the present value of taxes for the representative household. Here, in contrast, debt policy reshuffles the tax burden across generations.

[10]The effect on aggregate consumption in $t + 1$, $C_{t+1} = c_{0,t+1}N_t + c_{y,t+1}N_{t+1}$, is ambiguous because the tax cut will increase $c_{0,t+1}$ and lower $c_{y,t+1}$.

Policy implications

Real-world pension systems are not purely PAYG and many countries have accumulated pension funds. However, public pension savings are relatively small: the U.S. Social Security Administration is scheduled to deplete its trust fund by 2033.

Two factors have strained public pension systems in OECD countries. First, the number of retirees relative to the number of workers has increased and will continue to increase in coming decades. Population aging is driven by both lower mortality (retirees living longer) and by lower fertility. This can be interpreted as a lower n in the model above. Second, the productivity growth rate has fallen in recent decades (secular stagnation). For example, the U.S. growth rate of GDP per capita – γ in our model – fell from 2.3% between 1950 and 2000 to 1.2% between 2000 and 2020.

The analysis above suggests that pension promises are a form of debt. What is the total level of effective government debt for the U.S. federal government? A narrow definition of debt corresponds to the value of government bonds outstanding. For the U.S., federal debt held by the public was 94% of GDP in the second quarter of 2023.[a] A more comprehensive definition includes the implicit debt in the pension system, i.e., the present value of future federal pension promises. For the U.S. federal government this measure of debt is $65.9 trillion, or almost 2.5 times annual GDP.[b] This massive figure excludes the future costs of Medicare (the federal health care program for retirees). These two measures of government debt—94% versus 94+245=339%—are strikingly different. The *de facto* debt burden therefore depends on how seriously one should take promises about financial debt versus promises of future pension benefits. An outright default on nominal debt is ruled out by the U.S. Constitution. However, the government could generate a surprise increase in inflation and thereby inflate away some of the debt. Pension promises, in contrast, do not enjoy any constitutional protection and the government is always free to reduce social security benefits or raise the age at which people are eligible to collect them.

[a] Debt held by the public excludes the holdings of government debt by Federal government entities such as the Social Security Trust Fund, but includes debt held by the Federal Reserve.

[b] Source: 2023 OASDI Trustees Report, Table VI.F2.

15.7 Taxes and transfers as instruments for redistribution

An important function of government is to redistribute and provide social insurance. To this end, the tax and transfer system includes a wide array of taxes, social insurance programs and means-tested benefits at different levels of government (federal, state, and local). To illustrate the extent of redistribution embedded in the U.S. system, Figure 15.9 plots pre- versus post-government income for different quantiles of the pre-government income distribution. Pre-government income is income before taxes and transfers. Post-government income is disposable income, defined as pre-government income plus transfers minus taxes. The relationship is approximately linear, except at the lowest income percentiles. This suggests

that the U.S. tax- and transfer system can be well approximated by a log-linear function:

$$y - T(y) \quad = \quad \lambda y^{1-\tau}, \tag{15.25}$$

where y is pre-government household income, $T(y)$ is taxes minus transfers, and $y - T(y)$ is disposable income. The parameter λ controls the level of taxation, while the parameter τ can be interpreted as a measure of tax progressivity. To see this, note that when $0 < \tau < 1$, the tax system is *progressive* in the sense that the marginal tax rate $T'(y)$ is larger than the average tax rate $T(y)/y$ for any positive income level. Conversely, when $\tau < 0$, the marginal tax rate is lower than the average tax rate, $T'(y) < T(y)/y$, implying that taxes are *regressive*. When $\tau = 0$, the tax system is flat, with a constant marginal tax rate $T'(y) = T(y)/y = 1 - \lambda$.

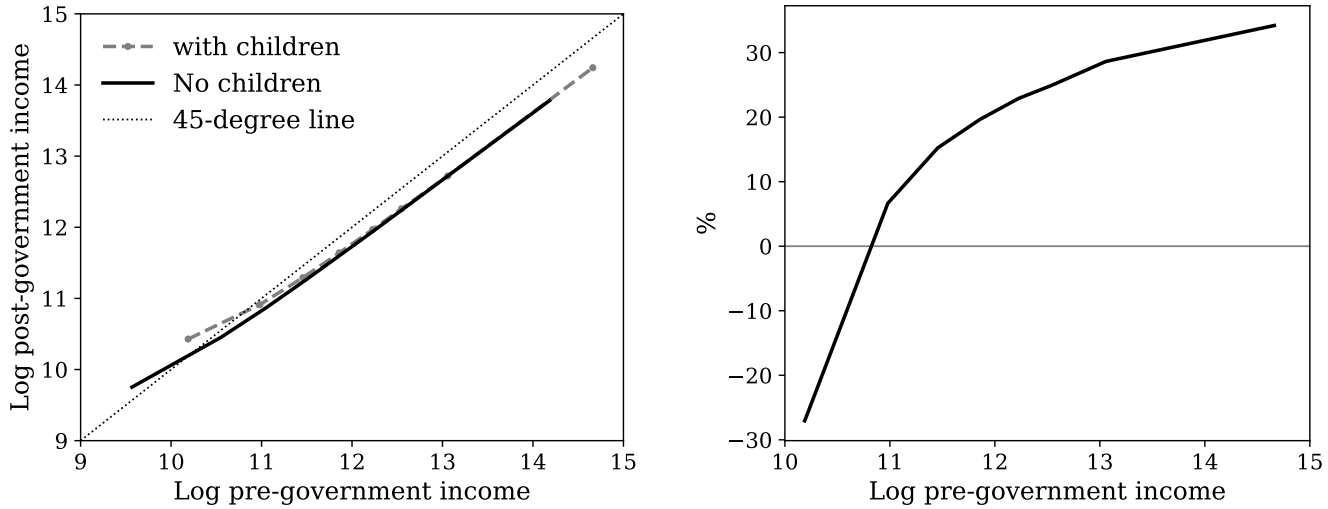


Figure 15.9: Left plot: Log of average post-government income plotted against log of average pre-government income for different quantiles of the distribution of pre-government income. Right plot: Average net tax rates by household income, defined as taxes minus transfers as a share of income, for households with children.

Source: Congressional Budget Office (CBO), 2016.

The right panel of Figure 15.9 plots average net tax rates, defined as taxes minus transfers divided by pre-government income. The picture illustrates that the average net tax rate is increasing with income. The U.S. tax and transfer system can therefore be said to be *progressive*.

15.7.1 A macro model of progressivity

We now illustrate the effects of tax progressivity on inequality and the macro economy using a simple static model of redistribution. The economy is populated by a unit continuum of individuals indexed by i. The utility function u is

$$u\left(c_i, \ell_i, G\right) = \log c_i - \frac{\ell_i^{1+1/\phi}}{1+1/\phi},$$

where c_i and ℓ_i are consumption and labor supply of individual i.

The tax and transfer system is assumed to take the log-linear form described in equation (15.25). The government's budget must be balanced, which imposes a constraint on the set of feasible fiscal policy choices (τ, λ, G).

The aggregate resource constraint dictates that output is spent on either private consumption or on public goods:

$$Y = \int_0^1 c_i \, di + G.$$

Individuals differ with respect to labor productivity. Their labor income is $w_i \ell_i$, where w_i is individual i's productivity. Individuals have no wealth, so consumption must equal disposable income:

$$c_i = \lambda \left(w_i \ell_i\right)^{1-\tau}.$$

Taking a first-order condition with respect to hours worked, one can solve in closed form for the equilibrium allocation. Hours worked and consumption are given by

$$\log \ell_i = \frac{\log(1-\tau)}{1+1/\phi} \tag{15.26}$$

and

$$\log c_i = \log \lambda + (1-\tau) \frac{\log(1-\tau)}{1+1/\phi} + (1-\tau) \log w_i. \tag{15.27}$$

Hours worked are falling in τ but are independent of the individual wage, w_i. Progressivity ($\tau > 0$) reduces hours because workers internalize that if they increase hours they will face a higher marginal tax rate, depressing the after-tax return. In the limit as $\tau \to 1$, workers anticipate that disposable income will equal λ, irrespective of hours worked, and thus hours will shrink to zero. Hours are independent of the wage because the utility function is in the balanced growth class.

Consumption is increasing in individual productivity, w_i. Tax progressivity dampens the pass-through from wages to consumption: a one percent increase in wages translates to a $1-\tau$ percent increase in consumption. Thus, tax progressivity reduces consumption inequality: the variance of log pre-government earnings is $\text{Var}(\log w)$, while the variance of log consumption is $(1-\tau)^2 \text{Var}(\log w)$. In conclusion, this simple economy illustrates the fundamental trade-off between efficiency and redistribution in a setting with an empirically plausible tax- and transfer system: higher progressivity reduces hours worked (lower efficiency) but also reduces consumption inequality (more redistribution). For more discussion of this trade-off, see Heathcote, Storesletten, and Violante (2017).

Chapter 16

Asset prices

Monika Piazzesi and Martin Schneider

16.1 Introduction

This chapter is about economies with multiple assets. The benchmark neoclassical growth model assumes that there is only one asset that agents can save in, capital, and hence only one rate of return. The model thus abstracts from differences in returns across types of capital, say business capital and housing. It further assumes that all capital is held directly: it is not "packaged" into other assets by capital owners or intermediaries. We do not distinguish, for example, firms' equity and debt, household mortgages and housing equity, mutual fund shares, pension accounts or bank deposits.

Allowing for multiple assets explicitly is interesting for two reasons. First, we can study more issues than in the benchmark model. For example, we would like to understand why rates of return differ on average or why some asset prices are more volatile than others. We can also ask why asset positions are different across agents. With multiple assets, agents decide not only on savings—the total value of assets held—but also make a portfolio choice—they pick the share of savings in each available asset. We can then ask how asset price fluctuations affect agents' welfare differently.

Second, allowing for multiple assets is typically important for conclusions we draw about allocations, or macroeconomic aggregates, even if we are not interested in asset prices or portfolios per se. One reason is that asset prices provide important moments to discipline assumptions on agents' objectives. For example, when we introduce equity and debt explicitly into the neoclassical growth model, we can infer what preferences of the representative agent are needed to match the average return on equity vs debt.[1] When markets are incomplete, there is another reason to study multiple assets: market structure matters directly for allocations. For example, agents bear less risk if there are instruments through which they can mutually insure each other.

[1] Asset prices are therefore relevant for macroeconomics even when markets are complete and allocations are efficient. Since complete markets allow us to find allocations from the planner problem and determine asset prices only in a second step, it is tempting to think that asset prices "don't matter". However, they are still important to test the model.

16.2 Background on household portfolio and asset prices

To motivate the study of multiple assets, we start with some background on household portfolios and asset price properties. Two common assets in household portfolios are housing and stocks. In the United States, most households (currently 65%) own a house, while half of households own stocks (currently 52%). These participation rates vary by country. For example, some countries like Germany have lower homeownership rates (47%), while other countries like Italy have higher homeownership rates (75%).

An important factor in determining the cross-country variation in stock market participation are pension systems. In the U.S., for example, workers choose how much of their income to save in pension plans and how to allocate these savings to various assets. Many households, therefore, participate in the stock market through these pension plans. In contrast, households in European countries often receive pension payments from the government, which are financed through taxes. In these countries, households often save less and do not actively face an asset allocation decision.

Within countries, there is much household heterogeneity in portfolio shares, the fraction of wealth invested in any given asset. Young households often do not have much wealth. Therefore, they borrow money from a bank to buy a house and thus hold a leveraged position in a risky asset. Older households accumulate wealth, slowly pay down their mortgage balance, and invest in both housing and stocks. There is also a substantial share of households with little savings at all.

Investing in long-lived assets can be risky. For example, stocks make future dividend payments that are random from today's point of view. Houses provide housing services that are random. For example, housing services can suddenly deteriorate in a natural disaster such as an earthquake. In addition to the randomness of assets' cash flows, the resale value of the asset is uncertain. The *payoff* of an asset in the next period, which consists of the cash flow and the resale value, can thus be either unexpectedly high or low.

In the data, the values of long-lived assets are highly volatile over time relative to their cash flows. Since both asset value and cash flows grow over time, it is helpful to study the ratio of asset price to cash flows. For stocks, we can measure the stock price at the end of the year and divide it by the dividends the stock paid during the year. Similarly, we can measure the house price at the end of the year and divide it by the value of housing services derived from the house during the year. For a currently rented house, the value of its housing services is simply the annual rent payment. For an owner-occupied house, the value of housing services is estimated based on data from comparable rentals.

Figure 16.1 shows the two ratios of asset price to cash flow in the United States for the years 1929-2023. The black line represents the ratio of stock values to the dividends paid per year by these stocks. The gray line represents an analogous ratio for housing, which divides the value of residential housing by the housing services it provides. These series show dramatic movements. Sometimes, these movements go together (such as in the Great Depression of the early 1930s) and sometimes in opposite directions (as in the Great Inflation of the late 1970s). Judging from these series, investing in stocks in 1999 was a bad idea, as they became significantly cheaper seven years later, while investing in housing was a good idea. How can we explain these movements, and how are they related to the rest of the economy?

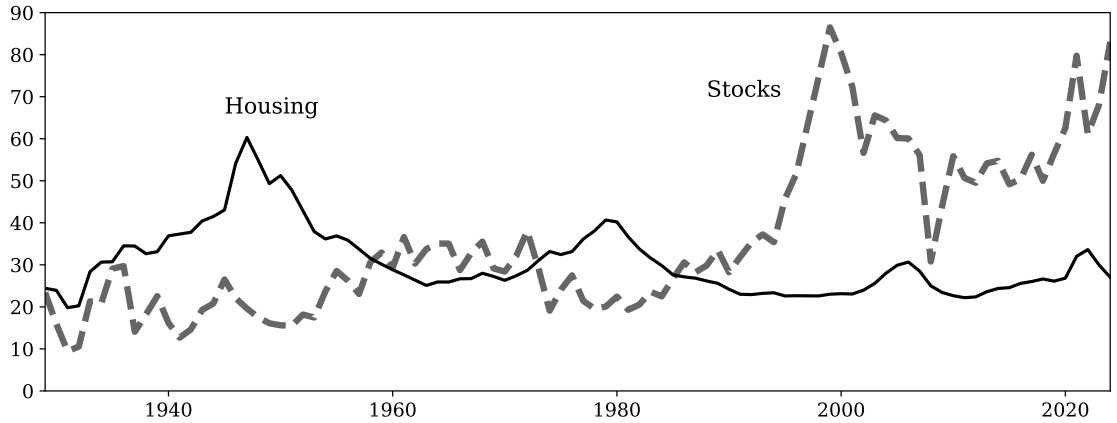


Figure 16.1: The ratio of asset value to cash flow for housing and stocks in the United States, 1930-2023.

16.3 Dynamic stochastic endowment economy

To understand the valuation of risky assets, we consider an endowment economy with an infinite horizon. At each date t, a state $\omega_t \in \Omega$ gets realized. There are S possible states. The history up to date t is $\omega^t = \{\omega_0, \omega_1, \ldots, \omega_t\}$ and has probability $\pi_t(\omega^t)$. We can connect the history at date t to its unique predecessor history ω^{t-1} by writing $\omega^t = \{\omega^{t-1}, \omega_t\}$. The conditional probability of the state ω_t is $\pi_t(\omega^t|\omega^{t-1})$ or, equivalently, $\pi_t(\omega_t|\omega^{t-1})$. There is a set $\mathcal{I}$ of agents so that the number of agents can be finite or infinite. These agents differ in their endowments of units of goods. Agent i has preferences described by a utility function $U_i(c_i)$ over consumption plans $c_i = \{c_{i,t}(\omega^t)\}_{t,\omega^t}$ and receives an endowment $y_i = \{y_{i,t}(\omega^t)\}_{t,\omega^t}$. The aggregate endowment is the sum of individual endowments $y_t(\omega^t) = \sum_{i\in\mathcal{I}} y_{i,t}(\omega^t)$.

An standard example for a utility function is expected utility

$$U_i(c_i) = \sum_{t=0}^{\infty} \sum_{\omega^t \in \Omega} \pi_t(\omega^t)\, \beta^t u\left(c_{i,t}(\omega^t)\right).$$

For understanding the dynamics of macro quantities, it is often sufficient to stay within this class of preferences. The most common example for the felicity function $u(x) = x^{1-\gamma}/(1-\gamma)$ is power utility, in which case the coefficient of relative risk aversion $\gamma = 1/\sigma$ is the inverse of the elasticity of intertemporal substitution σ (EIS). This specification thus tightly connects the agent's behavior towards gambles in static decisions (where γ matters) to the agent's behavior towards intertemporal choice in determistic settings (where σ matters.) For understanding the properties of asset prices, however, it will be important to derive results for more general utility functions. The leading example is Epstein-Zin utility, a utility function that separates risk aversion from the EIS, $\gamma \neq 1/\sigma$. This chapter will therefore derive results based on more general utility functions $U_i(c_i)$.

Planner problem A feasible allocation $\{c_{i,t}(\omega^t)\}_{t,\omega^t,i}$ of consumption across agents in this economy satisfies

$$\sum_{i\in\mathcal{I}} c_{i,t}(\omega^t) \leq \sum_{i\in\mathcal{I}} y_{i,t}(\omega^t) = y_t(\omega^t) \quad \text{for all } t, \omega^t.$$

The planner assigns a weight λ_i to each agent and selects a feasible allocation that maximizes the welfare function

$$U_\lambda(y) := \max_{\{c_i\}_{i\in\mathcal{I}}} \sum_{i\in\mathcal{I}} \lambda_i U_i(c_i) \quad \text{subject to } \sum_{i\in\mathcal{I}} c_{i,t}(\omega^t) \leq y_t(\omega^t).$$

The Lagrangian function of this problem is

$$\mathcal{L} = \sum_{i\in\mathcal{I}} \left\{ \lambda_i U_i(c_i) - \sum_{t=0}^{\infty} \sum_{\omega^t\in\Omega^t} \mu_t(\omega^t) \left(\sum_{i\in\mathcal{I}} c_{i,t}(\omega^t) - y_t(\omega^t) \right) \right\}.$$

The first-order condition (FOC) with respect to consumption of agent i at date t after history ω^t is

$$\lambda_i \partial U_i(c_i) / \partial c_{i,t}(\omega^t) = \mu_t(\omega^t). \tag{16.1}$$

An important property of this FOC is that every agent's weighted marginal utility is equated to a Lagrange multiplier that is independent of agent i. The envelope theorem says that the derivative of the Planner's Lagrangian function with respect to the aggregate endowment at date t after history ω^t is

$$\partial U_\lambda(y) / \partial y_t(\omega^t) = \mu_t(\omega^t). \tag{16.2}$$

If we combine the FOCs for agent i's consumption at different dates, we see that the planner equalizes the marginal rates of substitution across all agents i with the planner's MRS:

$$\frac{\partial U_i(c_i) / \partial c_{i,t+1}(\omega^{t+1})}{\partial U_i(c_i) / \partial c_{i,t}(\omega^t)} = \frac{\partial U_\lambda(y) / \partial y_{t+1}(\omega^{t+1})}{\partial U_\lambda(y) / \partial y_t(\omega^t)}. \tag{16.3}$$

Arrow-Debreu economy Suppose agents trade Arrow-Debreu securities at date 0, as in Section 7.4 of Chapter 7. These securities are claims to a unit of consumption at a specific date t after a specific history ω^t. The time-0 price of this security is the Arrow-Debreu price of consumption at date t after history ω^t denoted $p_t^0(\omega^t)$. Given consumption prices $p_t^0(\omega^t)$ and the endowment y_i of agent i, the budget set of agent i is

$$B^{AD}(p^0, y_i) = \left\{ \{c_{i,t}(\omega^t)\}_{t,\omega^t} : \sum_{t,\omega^t} p_t^0(\omega^t) c_{i,t}(\omega^t) \leq \sum_{t,\omega^t} p_t^0(\omega^t) y_{i,t}(\omega^t) \right\}.$$

Agent i selects the consumption plan in the budget set to maximize utility. The Lagrangian function of the agent's problem is

$$\mathcal{L} = U_i(c_i) - \alpha_i \left(\sum_{t=0}^{\infty} \sum_{\omega^t\in\Omega^t} p_t^0(\omega^t) c_{i,t}(\omega^t) - p_t^0(\omega^t) y_{i,t}(\omega^t) \right).$$

The first-order condition with respect to consumption at date t after history ω^t is

$$\partial U_i\left(c_i\right)/\partial c_{i,t}\left(\omega^t\right) = \alpha_i p_t^0\left(\omega^t\right). \tag{16.4}$$

As long as the budget constraint binds, the Lagrange multiplier α_i is strictly positive, and we can divide by $\alpha_i > 0$. We again obtain the key property that every agent's weighted marginal utility is equated to a variable, here the price of consumption, that is independent of agent i. By combining FOCs of agent i at different dates, we can see that trading at date 0 equalizes the marginal rates of substitution across agents:

$$\frac{\partial U_i\left(c_i\right)/\partial c_{i,t+1}\left(\omega^{t+1}\right)}{\partial U_i\left(c_i\right)/\partial c_{i,t}\left(\omega^t\right)} = \frac{p_{t+1}^0\left(\omega^{t+1}\right)}{p_t^0\left(\omega^t\right)}. \tag{16.5}$$

Arrow-Debreu equilibrium An equilibrium in the Arrow-Debreu economy consists of a consumption allocation $\{c_i^*\}_{i\in\mathcal{I}}$ and consumption prices p^{0*} such that (i) the consumption plan c_i^* is in the budget set $B^{AD}\left(p^{0*}, y_i\right)$ for given consumption prices p^{0*} and maximizes the utility of agent i, (ii) markets clear $\sum_{i\in\mathcal{I}} c_{i,t}\left(\omega^t\right) = y_t\left(\omega^t\right)$ for all dates t and histories ω^t.

First Welfare Theorem The optimality conditions in the Arrow-Debreu economy (16.4) and the planner problem (16.1) look similar. The planner wants marginal utilities of agents i and j to be collinear

$$\lambda_i \partial U_i\left(c_i\right)/\partial c_{i,t}\left(\omega^t\right) = \lambda_j \partial U_j\left(c_j\right)/\partial c_{j,t}\left(\omega^t\right) = \mu_t\left(\omega^t\right). \tag{16.6}$$

The optimization by individual agents given Arrow-Debreu prices $p_t^0\left(\omega^t\right)$ also leads to collinear marginal utilities

$$\frac{1}{\alpha_i}\partial U_i\left(c_i\right)/\partial c_{i,t}\left(\omega^t\right) = \frac{1}{\alpha_j}\partial U_j\left(c_j\right)/\partial c_{j,t}\left(\omega^t\right) = p_t^0\left(\omega^t\right). \tag{16.7}$$

If we select agent weights $\lambda_i = 1/\alpha_i$ in equation (16.6), we obtain the same conditions as in equation (16.7). The consumption allocations in the Arrow-Debreu equilibrium and the planner problem are thus the same. This is an implication of the First Welfare Theorem, which says that the consumption allocation in the Arrow-Debreu equilibrium is Pareto optimal. This property will ensure that marginal utilities of different agents are collinear and marginal rates of substitution across agents, as well as across planner and agents, are identical.

Sequential markets We now introduce repeated trading in consumption and asset markets, which involves assets like those we see being traded in actual asset markets instead of Arrow-Debreu securities. Suppose that agents can trade N assets and buy consumption in spot markets at every date t. Each asset n pays a dividend $d_t^n\left(\omega^t\right)$ in units of consumption at date t after history ω^t. The N-dimensional vector $d_t\left(\omega^t\right)$ contains the dividends of all N assets. The dividend stream $\{d_t\left(\omega^t\right)\}_{t,\omega^t}$ of the assets is exogenously given.

The N-dimensional vector of asset prices $p_t\left(\omega^t\right)$ describes the units of consumption needed to purchase the assets at date t after history ω^t. Our convention is that assets are traded

ex dividend: the buyer of the assets at date t will receive the dividends $d_{t+1}(\omega^{t+1})$ at date $t+1$ after history ω^{t+1}. The asset prices $p_t(\omega^t)$ are thus quoted *ex dividend*, which means without the dividend: the asset prices do not contain the dividends paid at date t.

Agents can buy or sell assets, starting from some initial holdings. Unless stated explicitly otherwise, we assume that the agents have no assets initially, $\theta^n_{-1} = 0$ for all n. If agents buy asset n, they hold a positive number $\theta^n_t(\omega^t) > 0$ of the asset. If they sell asset n, they short $\theta^n_t(\omega^t) < 0$ many assets. Alternatively, they can choose to not have any number of asset n, $\theta^n_t(\omega^t) = 0$. The N-dimensional vector $\theta_t(\omega^t)$ collects all asset holdings. We call $\theta = \{\theta_t(\omega^t)\}_{t,\omega^t}$ a *trading strategy*.

With sequential trading, the budget set of agent i is

$$B(p, y_i) = \Big\{ \{c_{i,t}(\omega^t)\}_{t,\omega^t} : \text{ there is a } \theta_i \text{ such that } c_{i,t}(\omega^t) + p_t(\omega^t)^\top \theta_{i,t}(\omega^t) \leq y_{i,t}(\omega^t) + \left(d_t(\omega^t) + p_t(\omega^t)\right)^\top \theta_{i,t-1}(\omega^{t-1}) \text{ for every } t, \omega^t \in \Omega^t \Big\}. \tag{16.8}$$

Spending on consumption and current asset holdings has to be less than the endowment and payoffs from previous asset holdings. These payoffs consists of dividends plus the current value of the assets. Agent i chooses a consumption plan $c_i \in B(p, y_i)$ given asset prices p and the endowment y_i to maximize utility $U_i(c_i)$.

Sequential market equilibrium A competitive equilibrium is a collection of consumption plans $c^*_i = \{c^*_{i,t}(\omega^t)\}_{t,\omega^t}$ and trading strategies $\theta^*_i = \{\theta^*_{i,t}(\omega^t)\}_{t,\omega^t}$ for each agent i and asset prices $p^* = \{p^*_t(\omega^t)\}_{t,\omega^t}$ such that (i) c^*_i is in the budget set $B(p^*, y_i)$ for given asset prices p^* and the agent's endowment y_i, and maximizes utility $U^i(c_i)$, (ii) good markets clear at every date after every history $\sum_{i\in\mathcal{I}} c^*_{i,t}(\omega^t) = \sum_{i\in\mathcal{I}} y^*_{i,t}(\omega^t)$, and (iii) asset markets clear $\sum_{i\in\mathcal{I}} \theta^*_{i,t}(\omega^t) = 0_{N\times 1}$ at every date after every history.

Here, we are assuming that assets are in *zero net supply*, the asset holdings across agents sum to zero. In some applications, we will want to assume that there is some exogenous supply $\bar{\theta}^j$ of asset j. For example, we may want to allow for a positive supply of government bonds. In this case, government bond market clearing becomes $\sum_{i\in\mathcal{I}} \theta^j_{i,t} = \bar{\theta}^j$.

Preview If agents have a rich set of assets to trade at every date t, markets are *complete*. Intuitively, agents with access to a rich set of assets can easily shift consumption units across time and states of the world. Since agents take asset prices as given, they trade assets at date t until they equate their valuation of the assets' payoffs at date $t+1$. Since payoffs are defined as baskets of consumption units delivered in the various states of the world ω_{t+1} next period, equating asset prices is equivalent to agreeing on the value of each future consumption unit, as long as agents have enough assets to trade. Formally, markets are complete if for any sequence $\{z_t(\omega^t)\}$, there exists a trading strategy $\theta = \{\theta_t(\omega^t)\}$ with payoffs $z_t(\omega^t) = \left(d_t(\omega^t) + p_t(\omega^t)\right)^\top \theta_{t-1}(\omega^{t-1}) - p_t(\omega^t)^\top \theta_t(\omega^t)$ for all histories and times $t > 0$.

With complete markets, we obtain three important results. First, the set of equilibrium allocations in the Arrow-Debreu economy and the sequential markets economy are the same. Second, the First Welfare Theorem holds, so that the equilibrium allocation is Pareto optimal. Third, there exists a representative agent. Without complete markets, all three results are lost.

The three results are key for understanding the degree to which agents' heterogeneity can matter for asset valuation. With complete markets, a single-agent model is sufficient for understanding asset prices because there is agreement among all agents on how to value consumption in each history. For heterogeneity to matter, markets have to be incomplete. With fewer assets to trade, agents will agree on the value of the assets they trade, at least in the absence of financial frictions (such as short-sale constraints). However, agents may disagree about the value of consumption in the various histories.

We will first gain intuition about these results in a version of this environment with only two dates. In the two-period economy, assets are traded at date 0 and not re-traded again at date 1. We will then study an infinite horizon economy in which consumption and assets are re-traded at every date.

16.4 Asset trading with two periods

Suppose there are S different states of the world that can happen tomorrow (at date 1). Each state ω has some probability $\pi(\omega) > 0$ with $\sum_{\omega\in\Omega} \pi(\omega) = 1$. Agent i receives an endowment $y_{i,0}$ units of consumption at date 0 and an endowment of $y_{i,1}(\omega)$ units of consumption in state ω at date 1. The agent has utility function $U_i(c_i)$ over consumption plans $c_i = \left\{c_{i,0}, \{c_{i,1}(\omega)\}_{\omega\in\Omega}\right\}$.

The agent can buy a portfolio θ of the N assets at date 0. We collect the N-dimensional payoffs $d(\omega)$ at date 1 in state $\omega \in \Omega$ in an $S \times N$ payoff matrix D for every state and asset tomorrow, as in Chapter 7 (Section 7.4). If there are Arrow securities for each state at date 1, then the payoff matrix is equal to the identity matrix, $D = I$. The N-dimensional vector p contains the asset prices at date 0, so p^n is the price of asset n in units of consumption at date 0. A portfolio θ involves spending $\sum_{n=1}^{N} p^n\theta^n = p^\top\theta$ units of consumption at date 0. At date 1, the agent receives portfolio payoffs $d(\omega)^\top\theta$ units of consumption in state ω at date 1. We collect the portfolio payoffs in an S-dimensional vector $D\theta$.

The agent maximizes utility

$$\max_{c_i\in B(p,y_i)} U_i(c_i) \tag{16.9}$$

where the budget set is

$$B(p,y_i) = \left\{ \left(c_{i,0}, \{c_{i,1}(\omega)\}_{\omega\in\Omega}\right) : \begin{array}{c} c_{i,0} + p^\top\theta_i \le y_{i,0} \\ c_{i,1}(\omega) \le y_{i,1}(\omega) + d(\omega)^\top\theta_i \end{array} \quad \text{for some } \theta_i \in \mathbb{R}^N \right\}. \tag{16.10}$$

Since the utility function is strictly increasing, the budget equations will bind, so they hold with equality. The budget equations for consumption in the various states ω at date 1 involve the portfolio payoff in each state.

Arbitrage We define a "free lunch," or arbitrage opportunity, as a portfolio that is either free today, $p^\top\theta \le 0$ and has non-negative payoffs tomorrow $D\theta \ge 0$ with one strictly positive payoff in at least one of the states, or a portfolio with a strictly negative value $p^\top\theta < 0$ that has nonnegative payoffs tomorrow $D\theta \ge 0$. An arbitrage is a strategy that would be infinitely

attractive for any agent who prefers more consumption over less. It makes sense to assume that such a strategy does not exist.

The compact way of defining an arbitrage opportunity is with the $(S+1) \times N$ matrix

$$W = \begin{bmatrix} -p^\top \\ D \end{bmatrix}, \tag{16.11}$$

which contains the payoff at date 0 and for every state at date 1. An arbitrage is a portfolio θ such that its payoff $W\theta \geq 0$ is nonnegative and strictly positive either today or tomorrow in at least one state of the world. Markets are arbitrage-free if $D\theta \geq 0$ implies $p^\top \theta \geq 0$ and $D\theta \geq 0$ with one strict inequality implies $p^\top \theta > 0$.

Set of attainable payoffs The set of attainable payoffs is

$$\mathcal{M} = \left\{ x \in \mathbb{R}^{S+1} : \text{ there is a portfolio } \theta \in \mathbb{R}^N \text{ such that } x = W\theta \right\},$$

where W is the matrix in equation (16.11). If the system of asset prices p allows for arbitrage opportunities, then the set $\mathcal{M}$ includes elements x that are non-negative and have at least one strictly positive entry.

State prices The price at date 0 of one unit of consumption in state ω at date 1 is called a state price. The contingent claim that will deliver one unit of consumption in state ω and zero in all other states is called an Arrow security.[2] A state price is thus an asset price, the price of an Arrow security. The absence of arbitrage implies that each state price $q(\omega)$ must be strictly positive. If the price of an Arrow security were zero or negative, one could construct a costless strategy—or even one that yields a gain today—that delivers a strictly positive payoff in state ω and zero otherwise, which constitutes an arbitrage opportunity.

Law of one price The law of one price says that two assets with identical payoffs in every state must trade at the same price. Again, if this was not true, then a strategy that buys the cheaper asset and sells the more expensive asset has zero payoffs at date 1 but makes a strictly positive payoff at date 0 — an arbitrage. For a set of state prices q, the law of one price therefore implies that the price of the nth asset is its future payoffs in the various states valued by the state prices.

$$p^n = \sum_{\omega \in \Omega} d^n(\omega)\, q(\omega). \tag{16.12}$$

By writing the key property of state prices in matrix form $p = D^\top q$, we can see that the vector of consumption prices

$$\hat{q} = \begin{pmatrix} 1 \\ q \end{pmatrix}, \tag{16.13}$$

[2]The Arrow-Debreu security that we introduced earlier is a contingent claim trading at date 0 for consumption at any date t, including $t > 1$, after history ω^t. An Arrow security is a contingent claim for *next period* consumption in one of the states of the world. Of course, the two concepts coincide in a model with only two periods. However, the statement that state prices are Arrow security prices will always be true.

which assigns a normalized price of 1 to consumption at date 0 and states prices q to consumption in the various states tomorrow, is orthogonal to the set of attainable payoffs $\mathcal{M}$. In other words, for any $x \in \mathcal{M}$, we have that $x^\top \hat{q} = 0$. This follows from (16.12) because for asset n we have $-p^n + \sum_{\omega \in \Omega} d^n(\omega) q(\omega) = 0$ and the attainable payoffs in $\mathcal{M}$ are simply linear combinations of the payoffs from trading the individual assets.

Fundamental theorem of asset pricing For given asset prices p, the absence of arbitrage is equivalent with the existence of state prices. In general, these state prices may not be unique, so we may have $p = D^\top q = D^\top q'$, where both q and q' are state prices. If markets are complete ($\text{rank}(D) = S$), state prices are unique. If the number of assets equals the number of states, $N = S$, the state prices can be computed by inverting the payoff matrix, $q = \left(D^\top\right)^{-1} p$. If there are more assets than states, $N > S$, there are redundant assets, and we can eliminate some of them to obtain a $S \times S$ payoff matrix D with rank S. The state prices can then again be computed by inverting the payoff matrix, $q = \left(D^\top\right)^{-1} p$. If $N < S$, the rank of the payoff matrix is not S and markets are incomplete. If this is the case, we may have many state prices.

Below we sketch the proof. For the equivalence of the two statements, we need to show that the existence of state prices implies no arbitrage. Suppose we have some state prices $q \in \mathbb{R}^S_{++}$ for which $p = D^\top q$, then the value of any portfolio $\theta \in \mathbb{R}^N$ satisfies $p^\top \theta = q^\top D\theta$. If $D\theta \geq 0$, then $p^\top \theta = q^\top D\theta \geq 0$ since state prices are strictly positive. If $D\theta > 0$, then $p^\top \theta = q^\top D\theta > 0$. Thus, there are no arbitrage opportunities given asset prices q.

The other direction of the proof–showing that no arbitrage implies the existence of state prices–is more work, but is instructive as to what state prices do. The key step in the proof separates the set of attainable payoffs $\mathcal{M}$ from the set of arbitrage opportunities $\mathcal{K} = \mathbb{R}^{S+1}_+$ as depicted in Figure 16.2 for the case of $S = 1$ tomorrow. The horizontal axis represents payoffs at date 0, while the vertical axis measures payoffs at date 1. To show that $\mathcal{M}$ and $\mathcal{K}$ only intersect at the origin, $\mathcal{M} \cap \mathcal{K} = \{0\}$, the proof uses a separating hyperplane theorem. The theorem implies that there is a nonzero linear function $F : \mathbb{R}^{S+1} \to \mathbb{R}$ such that $F(z) > 0$ for all nonzero $z \in \mathcal{K}$ and $F(x) = 0$ for all $x \in \mathcal{M}$.[3] This means there are some strictly positive cofficients $\tilde{q} = (\tilde{q}_0, \tilde{q}_1) \in \mathbb{R}^{1+S}_{++}$, that represent the nonzero linear function: $F(x_0, x_1) = \tilde{q}_0 x_0 + \tilde{q}_1^\top x_1$, where $x_0 \in \mathbb{R}$ and $x_1 \in \mathbb{R}^S$. The vector $\tilde{q}$, depicted as an arrow in Figure 16.2, is othogonal to the set $\mathcal{M}$, because if we evaluate the function at an element in $\mathcal{M}$, we get $-\tilde{q}_0 p^\top \theta + \tilde{q}_1^\top D\theta = 0$ for all $\theta \in \mathbb{R}^N$. The vector $\tilde{q}$ shown as an arrow in Figure 16.2 thus contains consumption prices (16.13), but does not necessarily assign a price of one to consumption at date 0. However, since $\tilde{q}_0 > 0$, we can normalize the price of consumption at date 0 on the horizontal axis to one, and find the state prices $q = \tilde{q}_1/\tilde{q}_0$ on the vertical axis.

The essence of this proof is the existence of a linear function $F : \mathbb{R}^{S+1} \to \mathbb{R}$ with strictly positive coefficients $\tilde{q}$ that values all assets and portfolios of assets. When we evaluate this

[3]The separating hyperplane theorem is for cones. A cone is a set X for which if $x \in X$ implies that $\lambda x \in X$ for any $\lambda > 0$. Both $\mathcal{M}$ and $\mathcal{K}$ are closed convex cones in $\mathbb{R}^{S+1}$ that intersect at the origin. The separating hyperplane theorem implies that there exists a nonzero linear function $F : \mathbb{R}^{S+1} \to \mathbb{R}$ such that $F(z) < F(x)$ for all z in $\mathcal{M}$ and $F(x) > 0$ for all nonzero $x \in \mathcal{K}$. However, since $\mathcal{M}$ is a linear space, $-z$ is also in $\mathcal{M}$. The inequality thus implies that $F(z) = 0$ for all z in $\mathcal{M}$ since F is nonzero. Otherwise we could always start with $F(z) < F(x)$ and get $F(-z) > F(x)$.

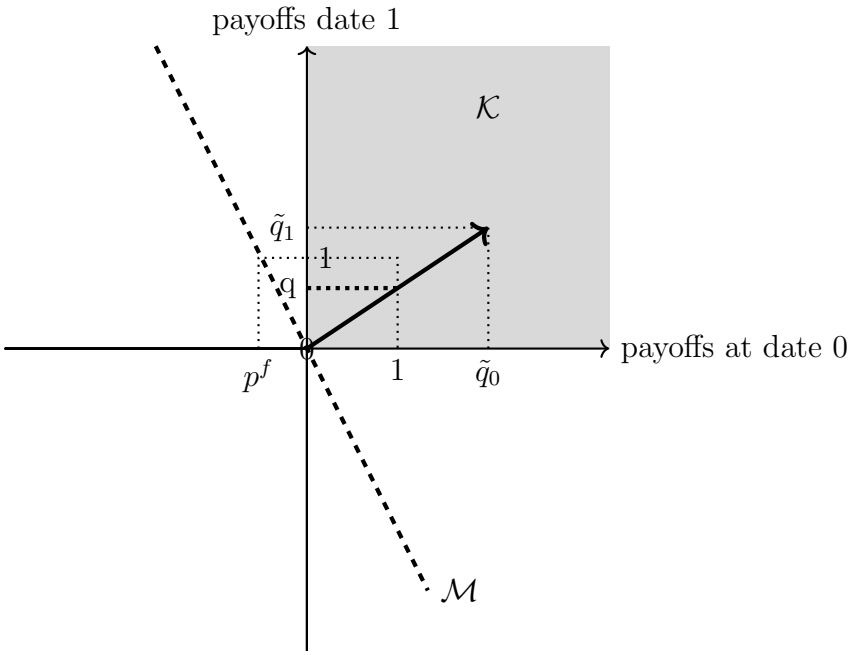


Figure 16.2: The set of attainable payoffs $\mathcal{M}$ and the set of arbitrages $\mathcal{K}$ separated by a linear function $F(x) = \tilde{q}_0 x_0 + \tilde{q}_1^\top x_1 = 0$ for $x \in \mathcal{M}$ with coefficients $\tilde{q}$ that are orthogonal to the elements of $\mathcal{M}$.

linear function for any attainable payoff, the function ties the date-0 value of any traded asset to the weighted sum of its date-1 payoffs, where the weights are state prices $q = \tilde{q}_1/\tilde{q}_0$. It is important to note that while the price p of any traded asset is given, the state prices q may not be unique. The theorem thus guarantees the existence of some state prices that justify the given prices p, but does not imply that the state prices themselves are unique, unless markets are complete. The following two examples illustrate this principle.

As an example, suppose there are two states, $S = 2$, and two assets: a stock and a bond that trade at given prices $p = \left(p^s, p^b\right)^\top$ and have a payoff matrix

$$D = \begin{pmatrix} \delta_1 & 1 \\ \delta_2 & 1 \end{pmatrix}, \qquad \delta_2 > 1 > \delta_1 > 0.$$

We impose two arbitrage restrictions. First, asset prices p have to be strictly positive, since the payoffs are strictly positive in both states. Second, we impose that the return on the stock cannot dominate that on the bond in all states (and vice versa): $\delta_1/p^s > 1/p^b > \delta_2/p^s$. The return is defined as the asset's payoff at date 1 divided by its price at date 0.

To determine state prices, we note that markets are complete markets, because $\text{rank}(D) = S = 2$. Moreover, we have as many states as assets $S = N$. In this case, we can recover unique state prices from asset prices by simply inverting the transpose of the payoff matrix:

$$q = \left(D^\top\right)^{-1} p = \frac{1}{\delta_1 - \delta_2} \begin{pmatrix} 1 & -\delta_2 \\ -1 & \delta_1 \end{pmatrix} \begin{pmatrix} p^s \\ p^b \end{pmatrix} = \frac{1}{\delta_1 - \delta_2} \begin{pmatrix} p^s - \delta_2 p^b \\ \delta_1 p^b - p^s \end{pmatrix}.$$

The state prices must be strictly positive, so we need the arbitrage conditions $\delta_1 p^b > p^s > \delta_2 p^b > 0$ as before.

The set of attainable payoffs tomorrow $D\theta$ is the $\mathbb{R}^2$. For every payoff $x \in \mathbb{R}^2$, there is a portfolio $\theta = D^{-1}x$ that will generate the payoff x. The value of the payoff x is the cost of that portfolio

$$p^\top \theta = p^\top D^{-1} x = \frac{1}{\delta_1 - \delta_2} \left(p^s - p^f \delta_2, p^f \delta_1 - p^s\right)^\top x.$$

Under the arbitrage conditions we imposed, the value of any nonnegative vector x with at least one strictly positive entry is strictly positive.

As another example, suppose again $S = 2$. However, we now only have one asset, a stock, with the same payoffs as before. D consists of a single column:

$$D = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}.$$

We have incomplete markets, because $\text{rank}(D) = 1 < S$. We are now restricted to payoffs $\begin{bmatrix} \delta_1 & \delta_2 \end{bmatrix}^\top \theta$ along a line in $\mathbb{R}^2$. Suppose $\delta_1 > \delta_2 > 0$. No arbitrage then requires that the stock price is positive. There are many $q \in \mathbb{R}^2_{++}$ so that $p^s = \sum_{\omega \in \Omega} d^s(\omega) q(\omega)$, where $d^s(\omega)$ is the payoff of the stock in state ω. Given our matrix D, we have $p^s = \delta_1 q(1) + \delta_2 q(2)$. For example, suppose $\delta_1 = 2$, $\delta_2 = 0.5$, and $p^s = 1$. Any $q(1) \in (0, 0.5)$ and $q(2) = 2 - 4q(1)$ are valid state prices, because they solve the single equation $p^s = 2q(1) + 0.5q(2)$.

Connection between no arbitrage and individual agent optimality If there is a solution to the individual agent problem, that implies that there is no arbitrage opportunities. Suppose there were abitrage opportunities available. In this case, agents can add a free lunch to their budget sets. Therefore, there cannot be an optimal solution since any candidate optimal portfolio θ^* can always be improved upon with another free lunch on top of it.

The converse statement also holds: If the utility function is continuous and there are no arbitrage opportunities, then there is a solution to the individual agent problem. The key here is to apply the fundamental theorem that implies that no arbitrage implies the existence of state prices to rewrite the budget set (16.10) using the state prices. This rewriting of the budget set helps to show that it is compact, so that the maximum theorem applies and guarantees the existence of a solution to the optimization problem.

Rewriting the budget constraint If there are no arbitrage opportunities, we have state prices q that satisfy $p = D^\top q$. We use these state prices to rewrite the budget equation for date 0 and obtain the net trade away from the endowment of agent i

$$c_{i,0} - y_{i,0} \leq -p^\top \theta_i = -q^\top D \theta_i = -\sum_{\omega \in \Omega} q(\omega) d(\omega)^\top \theta_i.$$

For the budget equation at date 1, agent i's net trade in state ω is

$$c_{i,1}(\omega) - y_{i,1}(\omega) \leq d(\omega)^\top \theta_i.$$

We can multiply the net trade in state ω by the state price $q(\omega)$ and add up all net trades at date 0 and the different states ω at date 1 . We then obtain a single budget constraint

$$c_{i,0} - y_{i,0} + \sum_{\omega \in \Omega} q(\omega) \left(c_{i,1}(\omega) - y_{i,1}(\omega)\right) \leq 0.$$

If markets are arbitrage-free, the budget set $B(p, y_i)$ in equation (16.10) can thus be rewritten as

$$B(p, y_i) = \Bigg\{ \left(c_{i,0}, \{c_{i,1}(\omega)\}_{\omega\in\Omega}\right) : \; c_{i,0} - y_{i,0} + \sum_{\omega\in\Omega} q(\omega)\left(c_{i,1}(\omega) - y_{i,1}(\omega)\right) \leq 0 \qquad (16.14)$$
$$\text{and } \{c_{i,1}(\omega) - y_{i,1}(\omega)\}_{\omega\in\Omega} \in \text{ span}(D)\Bigg\}.$$

The span restriction ensures that the agent's net trades can be achieved with the given assets in the economy. This restriction can reduce the budget set compared to the budget set in an Arrow-Debrew economy with consumption at date 0 as the numeraire and Arrow-Debreu prices q for consumption at date 1. If the span of the payoff matrix is a strict subspace of $\mathbb{R}^S$, the agent cannot freely obtain all consumption bundles even if they are affordable given the state prices q. The reason is that given the asset payoff matrix D, there may not be enough assets to achieve the net trades $c_{i,1}(\omega) - y_{i,1}(\omega)$ away from the endowment $y_{i,1}(\omega)$ that are required to get the desired bundle. Moreover, state prices are not unique.

As an example, suppose the agent does not care about consumption at date 0 and also has no endowment at date 0. The budget constraint for date 0 is $p^\top\theta \leq 0$ and for state ω at date 1 is $c_{i,1}(\omega) - y_{i,1}(\omega) \leq d(\omega)^\top \theta_i \leq 0$. With two states, $S = 1$, we can draw the budget set with complete markets as the light gray shaded area in Figure 16.3. The agent can reach any point below or on the dashed budget line going through the endowment $y_{i,1}$. With incomplete markets, the agent may be more constrained. For example, with only a riskless bond, the agent can add or subtract one unit of consumption in both states, and thus can only reach points on the thick black line with unit slope starting at the endowment point.

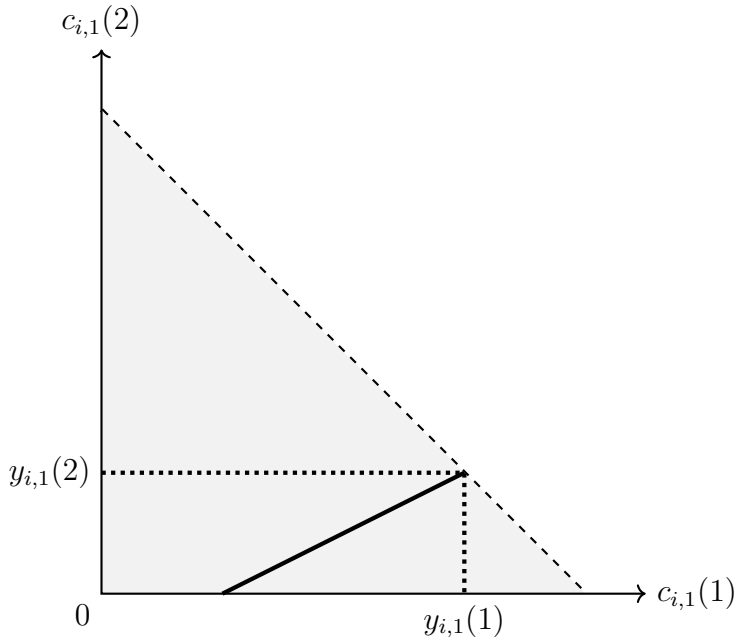


Figure 16.3: Budget sets with complete markets as dashed line and with only a riskless bond as solid line

Optimality conditions for asset holdings The Lagrangian for the single agent problem is

$$\mathcal{L} = U_i(c_i) - \psi_i\left(c_{i,0} + p^\top \theta_i - y_{i,0}\right) - \sum_{\omega\in\Omega} \eta_i(\omega)\left(c_{i,1}(\omega) - y_{i,1}(\omega) - d(\omega)^\top \theta_i\right),$$

where $\eta_i(\omega)$ is the Lagrange multiplier on the budget equation in state ω. The first-order condition for consumption at date 0 and in state ω at date 1 are

$$\frac{\partial U_i(c_i)}{\partial c_{i,0}} = \psi_i > 0 \text{ and } \frac{\partial U_i(c_i)}{\partial c_{i,1}(\omega)} = \eta_i(\omega) > 0.$$

The first-order condition for holdings θ_i^n of asset n is

$$-\psi_i p^n + \sum_{\omega\in\Omega} \eta_i(\omega)\, d^n(\omega) = 0.$$

Subjective state prices Combining, we obtain

$$p^n = \sum_{\omega\in\Omega} d^n(\omega)\, q^i(\omega), \text{ where } q_i(\omega) = \frac{\partial U_i(c_i)/\partial c_{i,1}(\omega)}{\partial U_i(c_i)/\partial c_{i,0}} \tag{16.15}$$

are state prices that satisfy the law of one price (16.12).

As an example, suppose agents have the same beliefs π and the same discount factor β. The utility function (16.9) takes the form

$$U_i(c_i) = \frac{c_{i,0}^{1-\gamma}}{1-\gamma} + \beta \sum_{\omega\in\Omega} \pi(\omega) \frac{c_{i,1}(\omega)^{1-\gamma}}{1-\gamma},$$

where γ is the coefficient of relative risk aversion, and $\pi(\omega)$ is the probability of state s. The resulting state prices are $q_i(\omega) = \beta\pi(\omega)\left(c_{i,1}(\omega)/c_{i,0}\right)^{-\gamma}$, where $c_{i,1}(\omega)/c_{i,0}$ is agent i's consumption growth rate from date 0 to state ω at date 1. A risk-averse agent is thus willing to pay a high price for consumption in state s when the consumption growth rate in that state is low. If the agent is risk neutral, $\gamma = 0$, the state prices $\beta\pi(\omega)$ just reflect the discounted probabilities of the states at date 1.

Disagreement about subjective state prices The agent index i is an important reminder that state prices are subjective. When some markets are missing, agents cannot freely trade consumption claims at date 0. The missing markets allow agents to disagree about the price $q_i(\omega)$ of a unit of consumption good in state ω at date 1: their marginal rates of substitution of consumption at date 0 for consumption in state ω at date 1 may be different. The subjective state price $q_i(\omega)$ of agent i may be different from the subjective state price $q_j(\omega)$ of agent j. In the example with expected power utility, the consumption growth rates may differ across agents. Moreover, agents can differ in their coefficients of relative risk aversion γ_i, discount factors β_i or beliefs π_i.

The optimality conditions for asset holdings lead to agreement among agents *about asset prices*. The first-order conditions (16.15) make sure that agents agree about the valuation of

the linear combination of payoffs. This highlights that assets promise bundles of consumption goods at date 1. By optimally trading these assets, agents will agree on the value of the bundle, but may disagree about the individual prices of consumption in the various states of the world. Going back to Example 2, agents i and j will agree about the stock price $p^s = 2q_i i\,(1) + 0.5q_i\,(2) = 2q_j\,(1) + 0.5q_j\,(2) = 1$, for example because they value consumption in each state according to the individual-specific state prices $q_i = (1/4, 1)^\top$ and $q_j = (1/3, 2/3)^\top$. With complete markets, there are enough assets with prices p^n that agents trade, that the first-order conditions (16.15) will lead them to agree on state prices. Put differently, at date 0 agents will assign the same value to extra units of consumption in the S states at date 1, $q_i = q_j$.

Agreement with complete markets If the span of the payoff matrix D is the entire $\mathbb{R}^S$, markets are complete. In this case, the span restriction in the budget set does not bind. The budget set with sequential trading is then identical to the budget set in an Arrow-Debreu economy with consumption at date 0 as the numeraire and state prices q for consumption at date 1. Moreover, the state prices q that determine the asset prices $p = D^\top q$ in the sequential trading economy are unique. The unique state prices solve $q = \left(D^\top\right)^{-1} p$ if $S = N$. If $N > S$, there are redundant assets, and we can compute unique state prices based on a subset of these assets that span the entire $\mathbb{R}^S$. Agents thus agree on the value of consumption in different states.

Stochastic discount factor If no arbitrage holds, there are state prices q, which we can use to define a stochastic discount factor $M(\omega) = q(\omega)/\pi(\omega)$. The price of asset n satisfies $p^n = \sum_{\omega\in\Omega} d^n(\omega)\, q(\omega) = \mathbb{E}[d^n(\omega)\, M(\omega)]$, while the return on the asset $R^n(\omega) = d^n(\omega)/p^n$ satisfies $1 = \mathbb{E}[R_n(\omega)\, M(\omega)]$.

The stochastic discount factor is a measure of how hungry the agent is for consumption. When $M(\omega)$ is high, the agent is hungry and desperate for more consumption. Any payoffs in those hungry "high M" states contribute a lot to the value of the asset, while payoffs in "low M" states do not contribute much. In the case of expected power utility, the stochastic discount factor is $M(\omega) = \beta\,(c_1(\omega)/c_0)^{-\gamma}$, which illustrates that the agent is hungry in states of the world ω in which the growth rate of consumption is low. If the agent is risk neutral, the stochastic discount factor is a constant, $M(\omega) = \beta$ in all states ω; the agent thus assigns equal value to consumption in the various states of the world.

Risk-neutral pricing A riskfree bond pays one unit of consumption in every state of the world at date 1. If agents are risk neutral, the bond price satisfies $p^b = 1/R^f = \mathbb{E}[M(\omega)] = \beta$. The price of any other asset n is given by its expected discounted payoffs

$$p^n = \mathbb{E}[d^n(\omega)\, M(\omega)] = \frac{\mathbb{E}[d^n(\omega)]}{R^f}, \tag{16.16}$$

where the discount rate is the riskfree interest rate.

Risk premium More generally, without assuming risk neutrality, the price of asset n satisfies

$$p^n = \frac{\mathbb{E}\left[d^n(\omega)\right]}{R^f} - \text{ risk adjustment},$$

where the risk adjustment reduces the price of the asset. What determines the risk adjustment? Using the definition of covariance,

$$\begin{aligned} p^n &= \mathbb{E}\left[d^n(\omega) M(\omega)\right] \\ &= \mathbb{E}\left[d^n(\omega)\right]\mathbb{E}\left[M(\omega)\right] + \operatorname{Cov}\left(d^n(\omega), M(\omega)\right) \\ &= \frac{\mathbb{E}\left[d^n(\omega)\right]}{R^f} - \left(-\operatorname{Cov}\left(d^n(\omega), M(\omega)\right)\right), \end{aligned}$$

we can see that the risk adjustment is minus the covariance of the asset's payoffs with the stochastic discount factor. The formula highlights that an insurance asset is particularly valuable: its payoffs are high precisely when the agent is hungry, the stochastic discount factor is high. An asset like a stock market index has payoffs that negatively covary with the stochastic discount factor: the stock market does well in economic booms, when consumption tends to be high, and badly in economic recessions, when consumption tends to be low. Compared to risk neutral pricing, this negative covariance is a positive risk adjustment which lowers the stock price.

An important insight is that not all risk matters for asset valuation. Only systematic risk, defined as risk that covaries with the stochastic discount factor, matters for asset values. Unsystematic risk does not affect asset values. Put differently, variance is not the appropriate measure of risk in financial markets, *covariance* is what matters.

Alternatively, we can rewrite the equation $1 = \mathbb{E}\left[R^n(\omega) M(\omega)\right]$ using the definition of covariance as

$$1 = \mathbb{E}\left[R^n(\omega) M(\omega)\right] + \operatorname{Cov}(R^n(\omega), M(\omega)) \Longleftrightarrow \mathbb{E}\left[R^n(\omega)\right] - R^f = -\frac{cov\left(R^n(\omega), M\right)}{\mathbb{E}\left[M\right]}. \tag{16.17}$$

The risk premium, defined as the expected excess return of the asset over the riskfree rate, is determined by minus the covariance of the asset's return with the pricing kernel (divided by the expected stochastic discount factor, which is a number close to one.) An insurance asset has positive covariance with the pricing kernel and therefore has a negative premium: the agent is willing to pay an insurance premium to get insurance. An asset like a stock has negative covariance with the pricing kernel and therefore has a positive equity premium: the agent has to be compensated with a high mean return to hold the asset.

Risk neutral probabilities We can define $\pi^*(\omega) = q(\omega) / \sum_{\omega' \in \Omega} q(\omega')$ and obtain probabilities that are stricly positive and sum up to one. The sum of state prices in the denominator of the probabilities, $\sum_{\omega' \in \Omega} q(\omega') = p^b = 1/R^f$, represent the value of one unit of consumption in every state next period, which is simply the price of a riskless bond. Therefore, asset prices are equal to the value of their payoff based on the state prices and, inserting $q(\omega) = \pi^*(\omega)/R^f$, we have

$$p^n = \sum_{\omega \in \Omega} d^n(\omega) q(\omega) = \frac{\sum\limits_{\omega \in \Omega} d^n(\omega) \pi^*(\omega)}{R^f} = \frac{\mathbb{E}^*\left[d^n(\omega)\right]}{R^f}, \tag{16.18}$$

where the expectation $\mathbb{E}^*$ is computed using probabilities π^*. The formula (16.18) looks like equation (16.16) which we obtained by assuming that agents are risk neutral. However, the probability measure π^* is not the actual probability measure π. Instead, it is a distorted probability measure under which asset prices are determined *as if* agents were risk-neutral. The probability measure π^* is therefore called risk-neutral probability. On Wall Street, financial economists often do not specify utility functions and determine equilibrium consumption streams to compute state prices or the pricing kernel. Instead, they back out the risk neutral probability π^* from observed asset prices and then use it, for example, to determine the price of derivatives. This chapter shows that state prices, pricing kernels, and risk-neutral probabilities are equivalent labels to talk about the determination of asset prices. Moreover, the risk-neutral probability π^* is unique if markets are complete, otherwise there may be many such probabilities, one for each possible state price.

Representative agent . With complete markets, the equilibrium asset prices of the heterogeneous-agent economy are identical to those in a representative-agent economy. More precisely, suppose a consumption allocation $\{c_i^*\}_{i\in\mathcal{I}}$ and portfolio holdings $\{\theta^{*i}\}_{i\in\mathcal{I}}$ together with asset prices p^* are a competitive equilibrium for the heterogeneous-agent economy with complete markets, then the consumption allocation $c = y$ and zero portfolio holdings together with the same asset prices p^* are a competitive equilibrium for the representative-agent economy with endowment $y = \sum_{i\in\mathcal{I}} y_i$ and preferences $U_\lambda \equiv \sum_{i\in\mathcal{I}} \lambda_i U_i$ where $\lambda^i = 1/\alpha_i$ and α_i is the Lagrange multiplier on the Arrow-Debreu budget constraint in the competitive equilibrium. The key to the proof of this result is to exploit the connection between the planner's problem and the decentralized equilibrium discussed in section 16.3. The decentralized equilibrium coincides with the solution to the planner's problem with the particular weights λ^i on the households. We can then interpret the planner's objective U_λ as the utility function of a representative agent.

With complete markets, we can read off the (unique) equilibrium state prices from the marginal rate of substitution of the representative agent. When we choose agent weights $\lambda^i = 1/\alpha^i$, combine the first-order conditions of individual agents (16.4) with the planner's first-order conditions (16.1) and envelope condition (16.2), we can write

$$q(\omega) = \frac{\partial U_\lambda(y)/\partial y_1(\omega)}{\partial U_\lambda(y)/\partial y_0}.$$

16.5 Dynamic asset trading

To define an arbitrage and to write the set of attainable payoffs, it will be useful to have a notation for the payoffs that can be generated by a particular trading strategy.

Payoffs generated by trading strategy The payoffs $\{x_t(\omega^t)\}_{t,\omega^t}$ generated by a particular trading strategy $\theta = \{\theta_t(\omega^t)\}$ are defined as

$$x_t(\omega^t) = \sum_{n=1}^{N} \left(p_t^n(\omega^t) + d_t^n(\omega^t)\right)\theta_{t-1}^n(\omega^{t-1}) - p_t^n(\omega^t)\,\theta_t^n(\omega^t).$$

With no initial asset holdings ($\theta_{-1}(\omega^0) = 0_{N\times 1}$), the initial payoff is minus the spending on the first portfolio, $x_0(\omega^0) = -p_0(\omega^0)^\top \theta_0(\omega^0)$. At date 1 after history ω^1, the initial portfolio $\theta_0(\omega^0)$ has payoff $(p_t(\omega^t) + d_t(\omega^t))^\top \theta_0(\omega^0)$ and the agent spends $p_1(\omega^1)^\top \theta_1(\omega^1)$ on the new portfolio, and so on.

Arbitrage Since agents trade at given asset prices p, an important question is whether assets offer a free lunch. An arbitrage is a trading strategy that delivers such a free lunch; these are units of consumption at some date and after some history for which agents don't have to pay anything. Formally, we define an arbitrage as a trading strategy θ that generates nonnegative payoffs $x_t(\omega^t) \geq 0$ for all dates t and histories $\omega^t \in \Omega^t$, with one strict inequality.

Set of attainable payoffs By forming a portfolio θ of assets, agents can obtain payoffs in the set

$$\mathcal{M} = \left\{ \left\{ x_t\left(\omega^t\right) \right\}_{t,\omega^t} : \text{there is a trading strategy } \theta \text{ that generates } x \right\}.$$

It will depend on the set of assets whether we can generate arbitrary payoffs x. We will say that *markets are complete* if any payoff stream $\{x_t(\omega^t)\}_{t>0,\omega^t}$ starting at date 1 can be generated at some cost at date 0. Otherwise markets are *incomplete*.

State prices The price at date t after history ω^t of consumption at date $t+1$ in state ω_{t+1} is denoted $q_{t+1}(\omega_{t+1}|\omega_t)$. This price is the value of a contingent claim that pays out one unit of consumption tomorrow, which is called Arrow security.

Law of one price, fundamental theorem, and individual optimality The properties we established in the two-period settings carry over to the dynamic trading environment. The law of one price implies that the price of asset n can be written as its payoffs valued by the state prices

$$p_t^n\left(\omega^t\right) = \sum_{\omega_{t+1}} \left(d_{t+1}^n\left(\omega^{t+1}\right) + p_{t+1}^n\left(\omega^{t+1}\right) \right) q_{t+1}\left(\omega_{t+1}|\omega^t\right).$$

The fundamental theorem of asset pricing states that no arbitrage is equivalent to the existence of state prices $q_{t+1}(\omega_{t+1}|\omega^t)$ that satisfy the previous equation. The proof of the theorem involves a separating hyperplane theorem that separates the set of attainable payoffs $\mathcal{M}$ from the set of arbitrage opportunities $\mathcal{K} = \{z_t(\omega^t) \text{ with } z_t(\omega^t) \geq 0\}$. Moreover, if there is a solution to the individual agent problem, that implies there are no arbitrages. Conversely, if the utility function is continuous and there are no arbitrages, there exists a solution to the individual agent problem.

Optimality conditions The Lagrangian for the single agent problem is

$$\begin{aligned}\mathcal{L} =& U_i(c_i) - \sum_{t,\omega^t} \eta_{i,t}\left(\omega^t\right) \Big(c_{i,t}\left(\omega^t\right) + p_t\left(\omega^t\right)^\top \theta_{i,t}\left(\omega^t\right) \\ & - y_{i,t}\left(\omega^t\right) - \left(d_t\left(\omega^t\right) + p_t\left(\omega^t\right) \right)^\top \theta_{i,t-1}\left(\omega^{t-1}\right) \Big).\end{aligned}$$

The first-order condition for consumption and holdings of asset n at date t after history ω^t are

$$\partial U_i\left(c_i\right)/\partial c_{i,t}\left(\omega^t\right)=\eta_{i,t}\left(\omega^t\right)$$

and

$$\eta_{i,t}\left(\omega^t\right)p_t^n\left(\omega^t\right)-\sum_{\omega^{t+1}}\eta_{i,t+1}\left(\omega^{t+1}\right)\left(d_{t+1}^n\left(\omega^{t+1}\right)+p_{t+1}^n\left(\omega^{t+1}\right)\right)=0.$$

Subjective state prices Combining, we generalize (16.15) to the multiperiod case

$$p_t^n\left(\omega^t\right)=\sum_{\omega_{t+1}}\left(d_{t+1}^n\left(\omega^{t+1}\right)+p_{t+1}^n\left(\omega^{t+1}\right)\right)q_{i,t+1}\left(\omega_{t+1}|\omega^t\right), \tag{16.19}$$

where

$$q_{i,t+1}\left(\omega_{t+1}|\omega^t\right)=\frac{\partial U_i\left(c_i\right)/\partial c_{i,t+1}\left(\omega^{t+1}\right)}{\partial U_i\left(c_i\right)/\partial c_{i,t}\left(\omega^t\right)}$$

and

$$M_{i,t+1}\left(\omega_{t+1}|\omega^t\right)=\frac{q_{i,t+1}\left(\omega_{t+1}|\omega^t\right)}{\pi_{t+1}\left(\omega_{t+1}|\omega^t\right)}.$$

Expected power utility Suppose the utility function takes the form

$$U_i\left(c_i\right)=\sum_{t,\omega^t}\pi_t\left(\omega^t\right)\beta^t\frac{c_{i,t}\left(\omega^t\right)^{1-\gamma}}{1-\gamma}. \tag{16.20}$$

The state price is

$$q_{i,t+1}\left(\omega_{t+1}|\omega_t\right)=\frac{\pi_{t+1}\left(\omega^{t+1}\right)\beta^{t+1}\left(\omega^{t+1}\right)c_{i,t+1}\left(\omega^t\right)^{-\gamma}}{\pi_t\left(\omega^t\right)\beta^t\left(\omega^t\right)c_{i,t}\left(\omega^t\right)^{-\gamma}}=\pi_{t+1}\left(\omega_{t+1}|\omega^t\right)\frac{\beta c_{i,t+1}\left(\omega^t\right)^{-\gamma}}{c_{i,t}\left(\omega^t\right)^{-\gamma}}. \tag{16.21}$$

Agreement or disagreement about state prices With complete markets, any payoff stream can be generated by trading the assets in the economy. When markets are complete, there are many equations (16.19) that determine asset prices as payoffs times the state prices. In particular, we can choose a full set of Arrow securities to complete markets and see that state prices, as well as the stochastic discount factor, will be unique. When markets are incomplete, there are fewer equations (16.19) than the number of state prices, and so state prices will not be unique.

Returns The return on asset n is defined as

$$R_{t+1}^n\left(\omega^{t+1}\right)=\frac{d_{t+1}^n\left(\omega^{t+1}\right)+p_{t+1}^n\left(\omega^{t+1}\right)}{p_t^n\left(\omega^t\right)}. \tag{16.22}$$

When M is a stochastic discount factor, the pricing equation (16.19) implies that the return satisfies the Euler equation

$$\mathbb{E}\left[M_{t+1}\left(\omega_{t+1}|\omega^t\right)R_{t+1}^n\left(\omega^{t+1}\right)|\omega^t\right]=1. \tag{16.23}$$

Using the definition of conditional covariance, writing the expectation conditional on history ω^t as $\mathbb{E}_t[.]$, and suppressing the dependence of returns and stochastic discount factor on history, we obtain

$$\mathbb{E}_t\left[M_{t+1}\right]\mathbb{E}_t\left[R^n_{t+1}\right]+\mathrm{Cov}_t\left(M_{t+1},R^n_{t+1}\right)=1,$$

implying

$$\mathbb{E}_t\left[R^n_{t+1}\right]-R^f_{t+1}=-\frac{\mathrm{Cov}_t\left(M_{t+1},R^n_{t+1}\right)}{\mathbb{E}_t\left[M_{t+1}\right]}.$$

The last equation determines the conditional expected excess return of asset n by the conditional covariance of its return with the stochastic discount factor. The difference to equation (16.17) is that the conditional covariance matters, which may vary over time. As a consequence, there will be times in which the conditional expected excess return is high, while it will be low at other times.

16.6 The equity premium puzzle and riskfree rate puzzle

We now want to understand the quantitative properties of equilibrium returns in a model with sequential markets trading. For this purpose, we make specific assumptions on preferences and endowments. In particular, we assume that there is a single agent with expected power utility (16.20) who receives an endowment stream y. We now add assets in zero net supply to this economy and solve for equilibrium asset prices such that the single agent chooses not to buy or sell these assets in equilibrium. Put differently, the agent chooses not to trade at given prices and just consumes the endowment.

More formally, a competitive equilibrium in this model consists of consumption c^*, a trading strategy θ^* and asset prices p^* such that (i) c^* solves the agent's optimization problem given asset prices p^*, (ii) goods markets clear $c^*=y$, and (iii) asset markets clear $\theta^*=0$. As long as markets are complete, the aggregation result says that the implications of this specification will be identical to those of a heterogeneous agent model in which agents have utility (16.20) and the aggregate endowment is y. An important lesson is that the quantitative implications of the single-agent model thus apply in much more general environments.

An equivalent formulation of this environment is the *Lucas tree model*. This formulation views stocks as trees that produce fruits every period. The fruits are the dividends that the owner of the trees receives. This is an interesting example, where some of the initial holdings of assets θ_{-1} are nonzero: in period 0, the single agent already owns the trees and can then buy or sell shares of the trees, which are in non-zero net supply. We normalize the supply of trees to one. There is no endowment. Given equilibrum prices p^*, the single agent chooses to hold the single tree $\theta^{s,*}=1$. Other assets, such as a one-period riskless bond, are in zero net supply and the single agent chooses to hold zero given p^*. In this formulation, equilibrium consumption equals the dividend $c^*=d$.

In the endowment economy, we assume that the endowment y has a growth rate that follows a two-state Markov chain with transition matrix Π. The equivalent Lucas tree model

formulation assumes that dividends d follow this Markov chain:

$$y_{t+1} = y_t G_{t+1}, \text{ where } G_{t+1} \in \{G_L, G_H\} \text{ with transition matrix } \Pi = \begin{pmatrix} \phi & 1-\phi \\ 1-\phi & \phi \end{pmatrix}.$$

The growth rate can either be low, G_L, or high, $G_H > G_L$. The transition matrix of the Markov chain is symmetric, so that the stationary probability π, which satisfies $\pi = \Pi^\top \pi$, spends half the time in each state, $\pi_1 = \pi_2 = 1/2$.

We want to think of the aggregate stock market as a claim to aggregate endowment, $d_t = y_t$, which is in zero net supply. The value of the stock market is therefore

$$p_t^s = \mathbb{E}_t \left[M_{t+1} \left(y_{t+1} + p_{t+1}^s \right) \right], \text{ with } M_{t+1} = \beta \left(c_{t+1}/c_t \right)^{-\gamma}. \tag{16.24}$$

In equilibrium, consumption equals the endowment, $c = y$, which means the growth rate of consumption equals $c_{t+1}/c_t = G_{t+1}$. Since the economy is growing, we solve for the stationary price-dividend ratio $v_t := p_t^s/y_t$

$$\frac{p_t^s}{y_t} = \mathbb{E}_t \left[M_{t+1} \frac{y_{t+1} \left(1 + p_{t+1}^s/y_{t+1} \right)}{y_t} \right] \quad \text{or} \quad v_t = \mathbb{E}_t \left[\beta G_{t+1}^{1-\gamma} \left(1 + v_{t+1} \right) \right].$$

Since we can substitute recursively for the price-dividend ratio v_{t+1} on the right-hand side, and the growth rate is a Markov process, the solution for v will be a function of only the current growth rate G_t.[4] Using the symbols of the Markov chain, and writing v_i for the price-dividend ratio in state i, we get

$$v_i = \sum_{j=1}^{2} \Pi_{ij} \beta G_j^{1-\gamma} \left(1 + v_j \right), \text{ or in matrix form } \mathbf{v} = \beta A \left(\mathbf{1} + \mathbf{v} \right),$$

where the matrix A is given by

$$A = \begin{bmatrix} \Pi_{11} G_1^{1-\gamma} & \Pi_{12} G_2^{1-\gamma} \\ \Pi_{21} G_1^{1-\gamma} & \Pi_{22} G_2^{1-\gamma} \end{bmatrix}.$$

We can therefore solve for the vector $\mathbf{v} = \left(I_{2\times 2} - \beta A \right)^{-1} \beta A \mathbf{1}$ by matrix inversion and obtain the price-dividend ratios in every state of the Markov chain.

We also want to solve for the (net) return on stocks

$$r_{t+1}^s = \frac{y_{t+1} + p_{t+1}^s - p_t^s}{p_t^s} = \frac{y_{t+1} \left(1 + v_{t+1} \right)}{p_t^s} - 1 = \frac{\left(y_{t+1}/y_t \right) \left(1 + v_{t+1} \right)}{v_t} - 1.$$

With our Markov chain, we want to solve

$$r_{ij}^s = g_j \frac{\left(1 + v_j \right)}{v_i} - 1 \tag{16.25}$$

[4] With recursive subsitution, we obtain

$$v_t = \mathbb{E}_t \left[\beta G_{t+1}^{1-\gamma} + \beta^2 G_{t+1}^{1-\gamma} G_{t+2}^{1-\gamma} + \beta^3 G_{t+1}^{1-\gamma} G_{t+2}^{1-\gamma} G_{t+3}^{1-\gamma} \ldots \right].$$

The conditional expected value of the return uses the transition probability Π_{ij} from state i to state j, while the unconditional expected value of the return uses the stationary probability π_i of state i

$$\mathbb{E}_i\left[r_{ij}^s\right] = \sum_{j=1}^{2} \Pi_{ij} r_{ij}^s \text{ and } \mathbb{E}\left[r_{ij}^s\right] = \sum_{i=1}^{2} \pi_i \mathbb{E}_i\left[r_{ij}^s\right]. \tag{16.26}$$

The price of a one-period riskless bond satisfies

$$p_t^f = \mathbb{E}_t\left[M_{t+1}\right] \text{ or } p_i^f = \sum_{j=1}^{2} \Pi_{ij} \beta G_j^{-\gamma}.$$

The real rate and its unconditional expected value are

$$r_i^f = \frac{1}{p_i^f} - 1 \text{ and } \mathbb{E}\left[r_i^f\right] = \sum_{i=1}^{2} \pi_i r_i^f. \tag{16.27}$$

The equity premium is defined as the difference between the expected stock return (16.26) and the mean riskfree rate (16.27). This unconditional moment can be compared to the average return difference in the data. One can also study a conditional version of the equity premium, defined as the difference between the conditional expected value of the stock return and the riskfree rate. This moment, which conditions on the available information at time t, can be compared to predictability regressions of future excess returns on stocks on regressors known at time t.

Our model has two preference parameters $\{\beta, \gamma\}$ and three parameters that describe the dynamics of the endowment process $\{G_L, G_H, \phi\}$. When we specify $G_L = 1 + \mu - \delta$ and $G_H = 1 + \mu + \delta$, the parameter μ is the mean consumption growth rate and δ generates variance in the consumption growth rate. The parameter ϕ is the probability of remaining in the current state of the Markov chain; it governs the persistence of consumption growth. We use the generalized method of moments to select the three parameters μ, δ and ϕ that describe the dynamics of the consumption growth rate. Table 16.1 shows three empirical moments that we want to match: average per-capita real consumption growth, its standard deviation (or volatility), and its autocorrelation. In the United States, consumption growth has a mean and volatility of roughly 2 percent, and a low autocorrelation of 10 percent.[5]

Table 16.1 shows the high average real return on the S&P 500, an index of 500 large companies which captures the behavior of the overall U.S. stock market. By returning 8 percent per year to investors, U.S. stocks have had a strong performance on average, especially compared to the 1 percent per year that investors would have achieved by holding a safe asset, the 3-month T-bill. The resulting (unconditional) equity premium of 7 percent

[5] We start from moments of the gross growth rate

$$\mathbb{E}\left[G_t\right] = \pi_1 G_L + \pi_2 G_H, \mathbb{E}\left[G_t^2\right] = \pi_1 G_L^2 + \pi_2 G_H^2, \mathbb{E}\left[G_t G_{t-1}\right] = \pi_1 \Pi_{11} G_L^2 + \pi_1 \Pi_{12} G_L G_H + \pi_2 \Pi_{21} G_H G_L + \pi_2 \Pi_{22} G_H^2.$$

The mean, variance and autocorrelation of the net growth rate are: $\mathbb{E}\left[G_t\right] - 1$, $\sigma^2(G_t - 1) = \sigma^2(G_t) = \mathbb{E}\left[G_t^2\right] - \mathbb{E}\left[G_t\right]^2$, and $\rho(G_t - 1, G_{t-1} - 1) = \rho(G_t, G_{t-1}) = \text{Cov}(G_t, G_{t-1})/\sigma^2(G_t) = \left(\mathbb{E}\left(G_t G_{t-1}\right) - \mathbb{E}\left(G_t\right)^2\right)/\sigma^2(G_t)$.

Table 16.1: Sample moments of aggregate consumption growth and real returns

	Consumption growth	Real return on S&P	riskfree rate
Mean	0.0175	0.0828	0.0074
Standard deviation	0.0233	0.1836	0.0380
Autocorrelation	0.1069		

Notes: Real per-capita data on nondurables and services consumption constructed as Fisher index based on annual sample 1929-2024 of NIPA Tables 2.3.5 in billions of Dollars, chain-type quantity indexes, 2.3.3., price indices 2.3.4., and population numbers in NIPA Table 2.1. Annual data from 1930 to 2024 on S&P returns are from Bob Shiller's "ie-data.xls" file at Yale. Annual data on the riskfree rate is from Bob Shiller's "chapt26.xlsx" file combined with data on the 3-month T-bill rate from 1954 to 2024 from FRED. To compute real returns and rates, we subtract the annual inflation rate based on nondurable and service consumption.

represents compensation for taking risk: the return on the S&P 500 has a volatility of 18 percent.

Once we estimate the dynamics of the endowment process, the only two free parameters are the discount factor β and risk aversion γ. Figure 16.4 computes the equity premium and riskfree rate for a discount factor β between 0 and 1, and a risk aversion γ between 0 and 100. The plot only shows equilibria in which the riskfree rate is below 4. The model is a qualitative success: the equity premium and the riskfree rate are both positive. However, the figure shows the massive quantitative failure of the model: the equity premium is less than 0.1 percent, while the riskfree rate quickly rises above its historical average of 1 percent. This tension between the model and the data is a puzzle for which Ed Prescott won the Nobel Prize in 2004. The tension is known as the equity premium puzzle and the riskfree rate puzzle.

16.7 Lognormal model

Computational results make it hard to understand the intuition behind the quantitative failure of the complete-markets, rational expectations model with power utility. To obtain analytical expressions, let us now assume that endowment growth is normally distributed:

$$g_{t+1} = \log(G_{t+1}) \sim N\left(\mu_g, \sigma_g^2\right) \tag{16.28}$$

With power utility, the log pricing kernel is lognormal as well, $\log M_{t+1} = \log \beta - \gamma g_{t+1}$. In this case, we can derive a closed-form expression for the riskfree rate. First, we determine the price of a one-period riskfree bond

$$p_t^f = \mathbb{E}_t\left[M_{t+1}\right] = \beta \mathbb{E}_t\left[e^{-\gamma g_{t+1}}\right] = \beta e^{-\gamma \mu_g + \frac{1}{2}\gamma^2 \sigma_g^2},$$

which implies that the riskfree rate is

$$r_t^f = \log\left(R_t^f\right) = \log\left(\frac{1}{p_t^f}\right) = -\log \beta + \gamma \mu_g - \frac{1}{2}\gamma^2 \sigma_g^2. \tag{16.29}$$

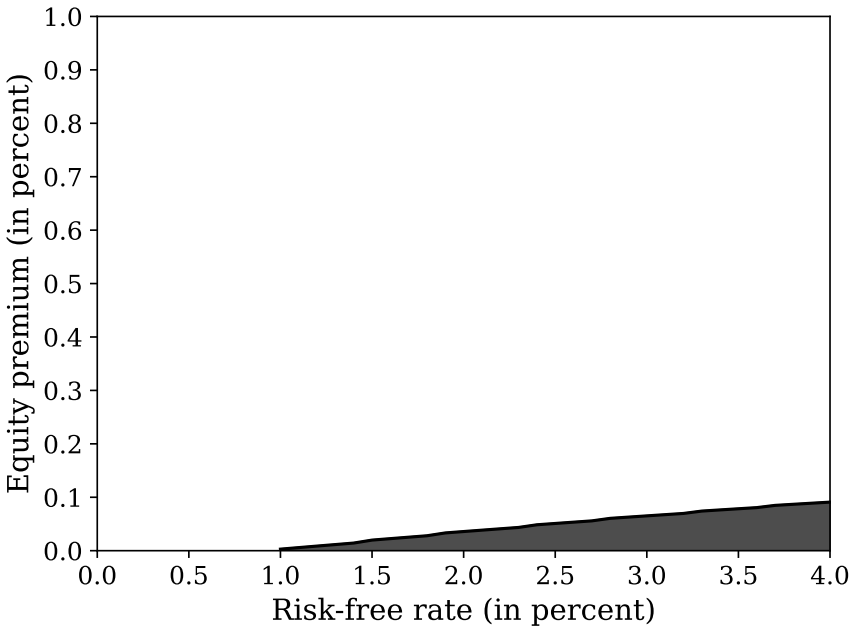


Figure 16.4: The shaded area shows the equity premium for equilibria in which the riskfree rate is below 4 percent for preference parameters $\beta \in [0, 1]$ and $\gamma \in [0, 100]$.

There are two key motives that determine the equilibrium riskfree rate. The *intertemporal smoothing* motive says that a higher expected consumption growth μ_g leads to a higher riskfree rate. The reason is that agents would like to smooth their consumption, try to borrow, and the riskfree rate increases to clear the bond market. The *precautionary savings* motive says that a higher volatility of consumption growth σ_g lowers the riskfree rate. To smooth consumption across states, agents try to save, and the riskfree rate decreases the clear the bond market.

With power utility, both motives are governed by the same parameter γ. With the sample moments from Table 16.1, $\mu_g = 1.75\%$ and $\sigma_g^2 = 0.054\%$, so that the intertemporal smoothing motive quantitatively dominates, except for extremely high values of risk aversion. In this case, higher risk aversion γ leads to a higher riskfree rate.

If the gross return on equity R^s and the pricing kernel are jointly lognormal, the Euler equation becomes

$$\begin{aligned} 1 &= \mathbb{E}_t\left[M_{t+1}R_{t+1}^s\right] = \mathbb{E}_t\left[\beta e^{-\gamma g_{t+1}+r_{t+1}^s}\right] \\ &= \beta \exp\left\{-\gamma\mu_g + \mathbb{E}_t\left(r_{t+1}^s\right) + \frac{1}{2}\left(\gamma^2\sigma_g^2 + \sigma_r^2 + 2\operatorname{Cov}_t\left(-\gamma g_{t+1}, r_{t+1}^s\right)\right)\right\} \end{aligned}$$

Taking logs and rearranging gives

$$\mathbb{E}_t\left(r_{t+1}^s\right) = -\log\beta + \gamma\mu_c - \frac{1}{2}\gamma^2\sigma_c^2 - \frac{1}{2}\sigma_r^2 + \gamma\operatorname{Cov}_t\left(g_{t+1}, r_{t+1}^s\right)$$

The first three terms in this expression are just the riskfree rate. Therefore, the equity premium is

$$\mathbb{E}_t\left(r_{t+1}^s\right) - r_t^f + \frac{1}{2}\sigma_r^2 = \gamma\operatorname{Cov}_t\left(g_{t+1}, r_{t+1}^s\right) \tag{16.30}$$

The term $\frac{1}{2}\sigma_r^2$ is a Jensen's inequality term which is introduced by studying log returns, since $\log \mathbb{E}_t\left(R_{t+1}^s\right) = \mathbb{E}_t\left(r_{t+1}^s\right) + \frac{1}{2}\sigma_t^2\left(r_{t+1}^s\right)$. The equity premium therefore includes the Jensen's term on the left-hand side.

The determination of the equity premium in equation (16.30) is the core of the *consumption-CAPM or CCAPM*. This name is derived from the CAPM (capital asset pricing model), which is derived with quadratic utility (or mean-variance preferences). According to the CAPM, assets that are highly exposed to overall stock-market risk earn higher returns on average. With power utility and lognormal consumption and (gross) returns, we obtain the CCAPM. According to the CCAPM, assets that are highly exposed to *consumption risk* earn higher returns. The equity premium is compensation for consumption risk exposures when holding stocks. The quantity of risk exposure is measured as the covariance of stock returns with consumption growth. The risk aversion coefficient γ represents the price for each unit of risk exposure. If investors are highly risk averse, they demand high compensation for being exposed to consumption risk.

We can write the covariance between consumption growth and stock returns as their correlation coefficient times the standard deviation of consumption growth σ_c and the standard deviation of stock returns σ_r. From Table 16.1, the product of these standard deviations is $\sigma_c\sigma_r = 0.0233 \times 0.1836 = 0.0043$. Even if consumption growth and stock returns are perfectly correlated, the right-hand side of equation (16.30) is, therefore, $\gamma \times 0.0043$. To obtain the 7% equity premium from Table 16.1, risk aversion would need to be high, $\gamma = 16$. With such a high risk aversion, the riskfree rate in equation (16.29) would be pushed to a high level, above its 1% mean in Table 16.1. This tension represents the joint equity premium and riskfree rate puzzle.

With the alternative assumption of Epstein-Zin utility (instead of power utility), we can reconcile a high equity premium with a low riskfree rate. The reason is that, with Epstein-Zin utility, the intertemporal smoothing motive in equation (16.29) is governed by the inverse of the elasticity of intertemporal substitution instead of risk aversion γ, while the parameter that enters the equity premium is still risk aversion. In this world, a high equity premium reflects that investors demand a high compensation for consumption risk, while they accept low compensation for holding riskfree bonds.

16.8 The excess volatility puzzle

Table 16.1 shows that consumption growth is only weakly autocorrelated. By assuming that consumption growth is i.i.d. lognormal, we obtain that the price-dividend ratio is constant:

$$v_t = \mathbb{E}_t\left[\beta G_{t+1}^{1-\gamma}\left(1+v_{t+1}\right)\right] = \mathbb{E}_t\left[\beta G_{t+1}^{1-\gamma} + \beta^2 G_{t+1}^{1-\gamma} G_{t+2}^{1-\gamma} + \beta^3 G_{t+1}^{1-\gamma} G_{t+2}^{1-\gamma} G_{t+3}^{1-\gamma} \ldots\right]. \quad (16.31)$$

The volatility of the price-dividend ratio determines the volatility of stock returns

$$r_{t+1}^s = \log\left(G_{t+1}\frac{\left(1+v_{t+1}\right)}{v_t}\right) = g_{t+1} + \log\frac{\left(1+v_{t+1}\right)}{v_t} \quad (16.32)$$

since the growth rate has a low volatility of 2 percent in Table 16.1. With a constant pd-ratio, the second term on the right-hand size of the equation (16.32) for stock returns is constant,

implying that the volatility of the return is also 2 percent. This low volatility stands in stark contrast to the high historical volatility of 18 percent in Table 16.1. The resulting tension between the model and the data is called the excess volatility puzzle for which Robert Shiller won the Nobel Prize in 2013.

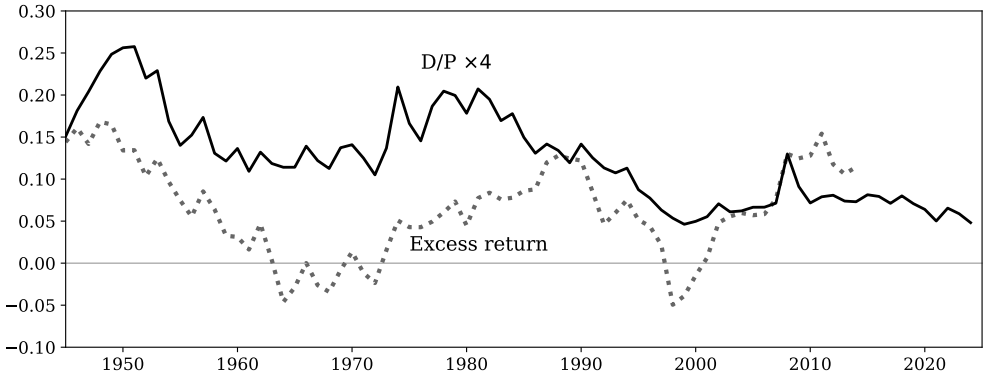


Figure 16.5: Dividend-price ratio together with excess returns on stocks over the next 10 years.

The preceding example assumes an iid growth rate g_{t+1}. Together with power utility, this assumption implies a constant price-dividend ratio and iid stock returns (16.32). Recent work on the quantitative asset-pricing implications of macro models has stressed that the price-dividend ratio varies dramatically over time. As a consequence, the second term in equation (16.32) is volatile and generates volatility in returns. Figure 16.5 shows annual postwar data on the inverse price-dividend ratio $1/v_t = d_t/p_t^s$, the so-called dividend yield, as a gray line together with (annualized) excess stock returns on the S&P 500 over the next decade in black. The dividend yield is multiplied by 4, so that the two series fit on the same vertical scale. The dividend yield moves around in a persistent way between 1% and 6.5%. This translates into persistent movements in the price-dividend ratio, ranging from 15 to 100. In particular, a high dividend yield tends to be associated with high subsequent excess returns, since the two series in Figure 16.5 move together.

Table 2 shows the results of predicting returns from buying the S&P 500 in year t and selling in year $t+k$. The prediction is based on the current dividend yield d_t/p_t^s. As suggested by Figure 16.5, the estimated slope coefficient in this regression is positive: a higher dividend yield predicts higher future excess returns. Table 16.2 shows that the slope-coefficient increases and becomes more significant over longer horizons k. The R^2 also increases with the horizon, meaning that excess returns can be forecasted more easily over longer holding periods, such as several years. Over shorter horizons, excess returns are harder to predict.

The literature has studied three main approaches to capture this predictability in macro models. The approach deviates from the iid assumption on growth rates. This can be done by introducing time-varying conditional moments in growth rates, which are then measured from historical data in a rational expectations approach. From equation (16.31), we can see that time-varying first moments $\mathbb{E}_t[G_{t+h}]$ for $h > 0$ will generate movements in the price-dividend ratio v_t. From equation (16.30), we can see that time-variation in conditional second moments $\text{Cov}_t\left(g_{t+1}, r_{t+1}^s\right)$ will translate into time-variation in the conditional expectation of

Table 16.2: Return forecasting regressions

Horizon k	b	$t(b)$	R^2
1 year	2.41	1.78	0.04
5 year	16.85	2.03	0.17
10 year	51.49	3.20	0.28

Notes:Note: The regression equation is $r^s_{t\to t+k} - r^f_{t\to t+k} = a + bd_t/p^s_t + \varepsilon_{t+k}$. The dependent variable is the excess return on the S&P 500 over the 3-month T-bill rate. Data are annual 1947-2024. The k-year regression t-statistic uses the Hansen and Hodrick (1980) correction when $k > 1$.

future excess returns. There is some empirical evidence supporting such time variation, but it is quantitatively not strong enough.

The second approach allows investors' subjective beliefs to feature more time variation in conditional moments of growth rates than what is measured in regressions. For example, these models assume sentiment, ambiguity-aversion or other deviations from rational expectations, which implies that the conditional expected value of growth rates varies over time. The empirical evidence in favor of these non-rational expectations comes from survey data on investors or CEOs.

The third approach takes the time-variation from preferences. For example, some models feature preference shocks that are heteroskedastic. These preference shocks are designed to match the evidence on volatility and predictability. Since these preference shocks are unobservable, the support for this modeling approach has to come from other implications of the model that match additional data (such as micro data on household portfolio choice) or that rationalize some observed policy actions by the government (such as the behavior of tax rates over the business cycle.)

Chapter 17

Money

Andreas Hornstein and Per Krusell

17.1 Introduction

So far, this textbook has exclusively discussed *real* variables. In the models covered, money is not present, and "dollars" have no special meaning. In microeconomics, on which the macroeconomic models here are based, the label money is sometimes used but merely to denote a numéraire. That is, the only prices that matter are relative prices between goods and services traded. In macroeconomics more broadly, however, money often takes center stage. First, people are concerned with *inflation* and the notion that it may erode the purchasing power of income. Second, one of the main macroeconomic policy tools available to governments is *monetary policy*, conducted by *central banks* by controlling the stock (supply) of money or the nominal interest rate, i.e., the rate of exchange between dollars at different points in time (overnight, or from year to year). Historically, monetary policy primarily meant money creation as a source of revenue (*seigniorage*): the printing of new money (notes and coins) is cheap and allows the government to help finance its operations. Today, the monetary policy conducted by central banks is more about maintaining price stability and limiting fluctuations in macroeconomic activity.

In most developed countries up until 2021, the need to control inflation from becoming too high had almost disappeared from public debate. Inflation rates had been maintained at well under 5 percent for three decades. Indeed, concerns often focussed on inflation being too low, with episodes nearing deflation during and around the Great Recession period. Then quite suddenly, following the coronavirus pandemic and Russia's 2022 invasion of Ukraine, inflation rates were well above 5 percent. Today, inflation control is again an important topic in developed economies.

Meanwhile, many less developed and emerging-market economies have struggled with inflation rates remaining at high levels, disrupting daily life in major ways. When inflation runs completely out of control, we speak of *hyperinflation*, which may seem merely like an intellectual curiosity to those who were fortunate enough not to have experienced one, but which quite clearly is disastrous for an economy. Thus, first-order questions for this chapter include: what determines the price level and causes inflation, how is human welfare affected by inflation, and how can stable prices be maintained?

These questions fundamentally hover around the "dollar bill" or *fiat money* more gener-

ally. Why does an intrinsically useless piece of paper have value and how do markets and government policy jointly determine its value? We will therefore introduce theories *of* money to explain the role money has in the economy. These theories satisfy Neil Wallace's *dictum* (see Wallace, 1998) that in a theory of money, the use of money should not be an assumption, but an outcome. We will also discuss less ambitious theories—they do not attempt to explain money's role—but are at least theories *with* money that are frequently used. Note that the classification of theories *of* or *with* money is distinct from the usual undergraduate textbook classification of money's three roles as a store of value, a medium of exchange, and a unit of account. With these frameworks, we will discuss the determination of inflation and discuss various options for monetary policy.

A particularly salient phenomenon in monetary economics is *indeterminacy*: the idea that there are many (competitive) equilibria in a given economic environment and that these equilibria may be associated with different real allocations. In a basic sense, this should perhaps not be surprising: money's value today ought to depend on its value in the future, which in turn depends on its value after that, and so on. Can hyperinflations, for example, simply be unfortunate equilibria in economies where there is also an equilibrium with a stationary price level? We will see that in many economic environments, the answer is yes. Thus, it seems that at least in theory, the view held by many, including Milton Friedman, that (hyper)inflation can only result as a consequence of the central bank allowing the money supply to increase (a lot), is not correct.[1] Whether hyperinflations are associated only with excessive money growth empirically is, however, not the focus of this chapter: the main purpose here is merely to cover basic monetary theory.

The need for theory with (or of) money in monetary economics is not a foregone conclusion, however: in the framework that is most frequently employed in policy-making settings today, the New Keynesian model developed in Chapter 18, money is not even present. That setting hence abstracts from monetary aggregates and instead focuses entirely on the role of nominal stickiness (of prices and wages) and how it can make monetary policy have significant real effects. In contrast, in the present chapter, prices will always be assumed to be flexible, so that the discussion can be focused sharply on more basic issues. However, we will demonstrate how the New Keynesian model without money can be motivated, namely, as the "cashless limit" of a model economy with money.

The chapter begins in Section 17.2 with the first fundamental model *of* money: the overlapping-generations model from Chapter 5, for which Samuelson (1958b) made the case that intrinsically useless paper money—*fiat* money—could have value under some circumstances. The idea here is that in the overlapping-generations setting, when the young's endowments and preferences are such that they want to save, in the absence of capital or other assets, there is no one they can lend to. However, the presence of fiat money can allow this saving. When fiat money has value, it is therefore an example of an asset bubble: money is an asset, without a dividend, and when people accept money in exchange for goods and services, it must be that people believe in its value only because others do. In this model, money functions as a *store of value*, and it has value so long as there is no other asset that

[1] Friedman (1963)'s famous speech in India includes the assertion "inflation is always and everywhere a monetary phenomenon." Friedman's message, which was based on empirical observation, was that money printing, i.e., central bank policy, is a necessary and sufficient condition for inflation to occur.

dominates money in return, e.g., bears interest. Moreover, the section demonstrates that there is equilibrium indeterminacy: apart from an equilibrium with valued money and a stable price level, there are also hyperinflationary equilibria as well as an equilibrium where money never has value.

Next, Section 17.3 looks at models with infinitely lived agents. There, we first show that money cannot have value, even when other assets are missing. The intuition behind this result is as follows. In a finite-horizon economy, money quite trivially cannot have value in the very last period, since at that time no seller of goods or services will accept money as payment. By induction, it is not valued in earlier periods either. Subsection 17.3.1 shows that this intuition survives also with an infinite horizon but, most importantly, the section sets up a general consumer budget constraint with money as well as government bonds. Next, we move to models *with* money, where a reduced-form liquidity value of money is assumed directly. Subsection 17.3.2 thus begins with the cash-in-advance model where money is assumed to be needed to purchase goods. The motivation is money's second function: the medium-of-exchange role. Similarly, in money-in-transactions-costs and money-in-the-utility-function models, real money balances are assumed to save on transactions costs and to give direct utility, respectively. These models feature rate-of-return dominance: other assets, that bear interest, do not yield these kinds of liquidity services (by assumption). The simple cash-in-advance model we present can also be seen as a motivation for the *quantity theory*: real balances are proportional to real output. Not only undergraduate textbooks but also some advanced research papers use a demand for money formulation that is simply an assumed quantity equation.[2]

Using the reduced-form models, in Subsection 17.3.3 we then briefly discuss optimal monetary policy (absent any stabilization concerns), including the well-known *Friedman rule*. We also show that models incorporating reduced-form liquidity can likewise exhibit equilibrium indeterminacy. We then move to several other important conceptual topics. One is the cashless limit mentioned above: the idea that it is possible to use nominal quantities without even having a money stock present in the model. One of the main purposes of this subsection is to prepare the ground for Chapter 18 on New Keynesian models, where we assume that money is absent but that prices are sticky in nominal units. The cashless limit is non-trivial and involves the notion of *monetary policy rules*, such as interest-rate rules. In particular, when the interest rate is set as an increasing function of the price level, a range of equilibria are eliminated. This insight underlies the use of *Taylor rules* in New Keynesian models and in practical policy-making, where interest rates respond to the price level or the inflation rate. Finally, yet another policy rule that has received significant attention involves the interaction between fiscal and monetary policy: the *fiscal theory of the price level*. We explain its origin and the intuition behind why it can be seen as a mechanism for eliminating equilibria and, yet, is controversial.

We then look at multiple currencies, that is, exchange rates, in Section 17.4. The idea is not to venture into international economics; rather, we discuss the implications of our basic theories of money for the relative values of different fiat currencies that might even circulate within a given economy. This section therefore also involves a short discussion of crypto-currency.

[2]See, e.g., Mankiw and Reis (2002).

In the remaining two sections of this chapter we briefly discuss two fundamental theories of money for which valued money is not an assumption but an outcome of the particular environment. The first approach builds on money as a store of value in an environment with infinitely-lived agents and a limited set of other assets, Section 17.5. In Section 17.6, we look at theories *of* money as a medium of exchange. Here, the idea is that search frictions among traders of goods/services with an *absence of double coincidence of wants* can be seen as a deep friction motivating a value for fiat money.

17.2 Money in overlapping-generation models

Historically, various kinds of currencies have circulated, often in the form of real objects of intrinsic value (such as precious metals), with a variety of price stability and longevity outcomes. However, today money—as defined by notes and coins—is fiat, i.e., it has no intrinsic value, and it is not "backed" (say, by gold). Rather, people voluntarily choose to accept money as a means of payment for goods and services because they expect money to have real value later when they want to use it. Money is thus an asset, but not one that promises anything real of direct value. Asset pricing, as discussed in Chapter 15, should thus be useful, but we need to specify more clearly what the potential future benefits of money might be for someone considering accepting it today.

In this section, we look at money as a store of value. For that, we begin with the overlapping-generations model, which already Samuelson (1958b) realized could be used to explain why money—under some conditions—will be valued in equilibrium, despite being fiat. Thus, valued money is an outcome, not an assumption. In fact, there is another plausible outcome where money has no value because people do not believe it will ever have value. Let us now revisit the overlapping-generations model considered in Chapter 5. We proceed using simple examples, which can be easily elaborated on and extended.

17.2.1 An endowment economy

So, first, consider an endowment economy without production or storage. Each period a new cohort enters that lives for two periods and there is a representative agent per cohort. The endowment vector of any agent entering the economy at time 0 or later is (ω_y, ω_o), and the agent's preferences are assumed to be logarithmic: $\log c_y + \log c_o$. The initial old representative agent at time 0 has endowment ω_o and utility that is strictly increasing in c_o. Thus, the environment is stationary.

There is fiat money in the environment: an amount M of perfectly divisible and intrinsically useless objects. We assume that the initial old ones own these. The core question now is whether fiat money can have value. The idea here is that money can be used as a *store of value*. As such, it may be valuable since the endowment economy does not allow saving (and the young have no one to lend to who will be able to pay back).

Let $p_{m,t}$ denote the time t value of a unit of money in terms of consumption goods at time t. Thus, money has value if $p_{m,t} > 0$. But if $p_{m,t} = 0$ for all t, we have a *non-monetary* equilibrium, where no one values money. Then, the maximization problem of the generation

t agent is

$$\max_{c_y, c_o, M'} \log c_y + \log c_o \tag{17.1}$$

subject to

$$c_y + p_{m,t} M' = \omega_y, \quad c_o = \omega_o + p_{m,t+1} M', \quad \text{and} \quad M' \geq 0.$$

The last inequality is natural: money can only be held in positive amounts. This is unlike other assets we discussed so far; for example, we assumed that one can borrow by issuing bonds, that is, holding negative amounts of bonds. The agent of generation -1 makes a trivial decision and simply sets $c_{o,0} = \omega_o + p_{m,0} M$. Recall that the initial old agents begin with the entire stock of money M.

We will continue to assume, for now, that money has value, that is, $p_{m,t} > 0$. We can then combine the constraints from (17.1) to describe the available budget set for the consumer, without explicitly involving money. This delivers

$$c_y + \frac{c_o}{p_{m,t+1}/p_{m,t}} = \omega_y + \frac{\omega_o}{p_{m,t+1}/p_{m,t}} \quad \text{and} \quad \omega_y - c_y \geq 0. \tag{17.2}$$

The implied budget set is presented in Figure 17.1; the inequality constraint, which reflects non-negative money holdings, means that consumption when young cannot exceed endowments when young. As can be seen, the gross real return on money, $p_{m,t+1}/p_{m,t}$, is the slope of the budget constraint.

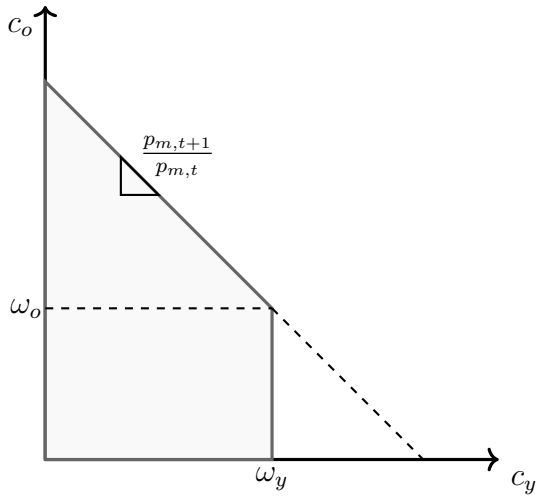


Figure 17.1: Budget set in the economy with fiat money

Solving the maximization problem is straightforward. Since the consumer's indifference curves are strictly convex and strictly decreasing we realize from the figure that either the solution is a point on the downward-sloping budget line that is tangent to the indifference curve—and, hence, the inequality constraint is slack—or it is right at the kink of the budget set, with $c_y = \omega_y$.[3] To find out which case applies, first ignore $\omega_y - c_y \geq 0$ and solve the

[3] Alternatively, use Kuhn-Tucker optimization; here, the Kuhn-Tucker multiplier will be zero in the former case and positive in the latter.

first-order conditions combined with the budget. The solution becomes

$$c_y = \frac{1}{2}\left(\omega_y + \omega_o \frac{p_{m,t}}{p_{m,t+1}}\right)$$

and

$$c_o = \frac{1}{2}\left(\omega_y + \omega_o \frac{p_{m,t}}{p_{m,t+1}}\right)\frac{p_{m,t+1}}{p_{m,t}}.$$

Now check to make sure that $c_y \leq \omega_y$. This amounts to

$$\omega_y \geq \omega_o \frac{p_{m,t}}{p_{m,t+1}} \quad \Longleftrightarrow \quad \frac{p_{m,t+1}}{p_{m,t}} \geq \frac{\omega_o}{\omega_y}.$$

That is, if the gross real return on money is at ω_o/ω_y or above, the solution is valid. The smaller is the ratio ω_o/ω_y, the larger is the consumer's desire to smooth consumption over time and save, but of course, the consumer is also dissuaded from saving if the return on money is low. If $p_{m,t+1}/p_{mt} < \omega_o/\omega_y$, the solution is $c_y = \omega_y$ and $c_o = \omega_o$ (and $M' = 0$): the consumer would want to buy a negative amount of money (borrow), but cannot. In sum, individual money demand by the young at t as a function of the money prices at t and $t+1$ is

$$p_{m,t} M'_t = \max\left\{\frac{1}{2}\left(\omega_y - \frac{\omega_o}{p_{m,t+1}/p_{m,t}}\right), 0\right\}. \tag{17.3}$$

Given the demand function, we can solve for equilibrium, which amounts to the young buying the entire money stock from the old:

$$M'_t = M \;\forall\; t. \tag{17.4}$$

There are two cases to consider. In one, money has no value at any point in time. This amounts to $p_{m,t} = 0$ for all t.[4] Recall that, in this case, (17.2) is not valid. Instead, the consumer is simply unable to save because no matter how much money they acquire, it will be worthless when they are old. Hence, there is an equilibrium where money does not have value.

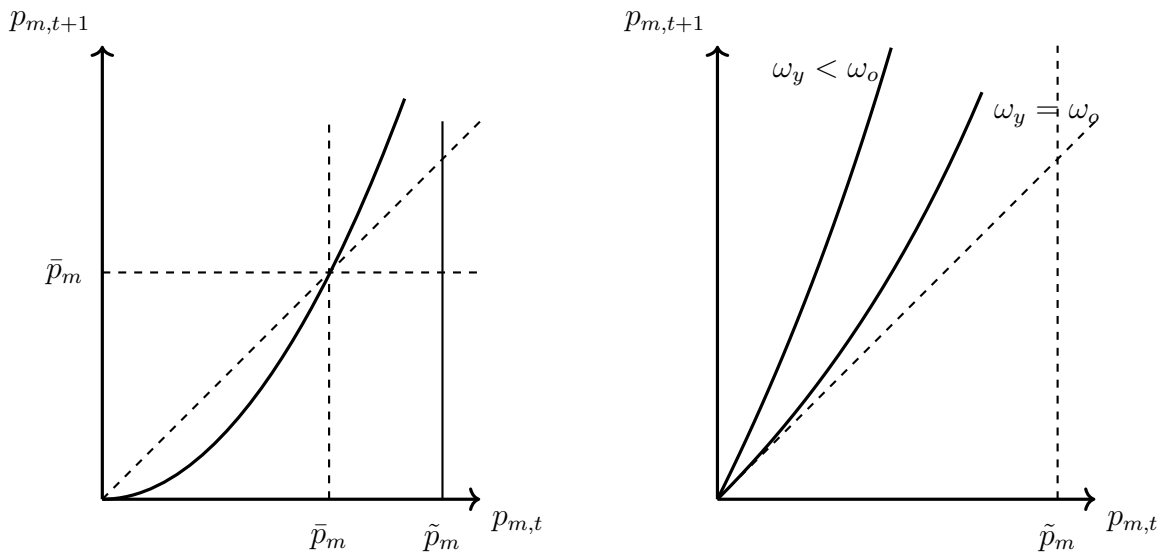
If $p_{m,t}$ is instead positive so that money has real value, then we can combine the expression for money demand (17.3) with the money market clearing condition (17.4) and solve for next period's price of money as a function of the current period's price,

$$p_{m,t+1} = \frac{\omega_o p_{m,t}}{\omega_y - 2M p_{m,t}}. \tag{17.5}$$

This is a non-linear first-order difference equation where the initial value $p_{m,0}$ is not given: it is endogenous (this is the whole point!). Thus, if we can find a solution to this difference equation where $p_{m,t}$ is positive at all points in time, we have a monetary equilibrium. For this, consider the following two cases: (i) $\omega_y > \omega_o$, and (ii) $\omega_y \leq \omega_o$. First, we can see that there is a constant solution to this difference equation: $p_{m,t} = \bar{p}_m = (\omega_y - \omega_o)/(2M)$. This value is positive for case (i), but it is not for case (ii).

Is the constant solution of case (i) the only possible solution or could there be non-stationary equilibria, where the price level changes over time? We plot the dynamics of equation (17.5) for cases (i) and (ii) in the two panels of Figure 17.2.

[4] In this case, $M'_t = M$ can still be assumed to hold: if money is free (and of no use), we can assume it is passed on in full from generation to generation.



(i) $\omega_y > \omega_o$ (ii) $\omega_y \leq \omega_o$

Figure 17.2: Price Dynamics

In Figure 17.2 (i) we plot the difference equation for case (i). The difference equation is defined only for $p_{m,t} < \tilde{p}_m = \omega_y/2M$, since for any current price equal or greater than $\tilde{p}_m$ next period's price is infinite or negative. On the interval $[0, \tilde{p}_m)$, the difference equation is monotonically increasing and it has two steady states at $p_{m,t} = 0$ and $p_{m,t} = \bar{p}_m$. The slope of the difference equation at the two steady states is less than one if money has no value and greater than one if money has value. Now consider an initial price for money $p_{m,0} > \bar{p}_m$. Following the kind of Solow-picture dynamics from Chapter 3, no such value can be part of an equilibrium: sooner or later, $p_{m,t+1}$ will exceed $\tilde{p}_m$, which cannot be an equilibrium. On the other hand, if $p_{m,0} < \bar{p}_m$, we see that the price dynamics will involve monotonically decreasing prices that converge to zero. Thus, money will be steadily losing value and have no value in the limit.

Case (ii), on the other hand, is one in which no monetary equilibrium exists. We plot the difference equation for case (ii) in Figure 17.2 (ii), and it is again increasing on the interval $[0, \tilde{p}_m)$. But now there is only one steady state at the origin when money has no value. Furthermore, the slope of the difference equation at the origin is equal (greater) one when $\omega_o = \omega_y$ ($\omega_o > \omega_y$) and increases in the current price. Thus, for any positive $p_{m,0}$, $p_{m,t+1}$ will eventually exceed $\tilde{p}_m$, although how long this takes depends on how high a $p_{m,0}$ is selected.

To summarize. If $\omega_y > \omega_o$, there is a continuum of equilibria indexed by an initial price level $p_{m,0} \in [0, \bar{p}_m]$. Of these different initial values of money, only one, $p_{m,0} = \bar{p}_m$, gives money a positive value in the limit, and for all the others, the value of money goes to zero. And if $\omega_y \leq \omega_o$, the only equilibrium is one where money has no value ever, $p_{m,t} = 0$.

Finally, note that the number of nominal units in the economy does not have a real importance: It is possible to rewrite the equilibrium difference equation (17.5) as $m_{t+1} = \omega_o m_t / (\omega_y - 2m_t)$, where $m_t \equiv p_{mt} M$. Thus, the equilibrium determines the real value of the total money stock, independently of how many fiat money units are available.

17.2.2 Welfare comparisons across equilibria

In this model, money derives its value from its role as a store of value, which would otherwise be missing. By fulfilling this function, money can help agents smooth consumption and provides a welfare benefit.

Recall the welfare characterization of overlapping-generations equilibria from Chapter 5: the Pareto efficiency of equilibria depends on the asymptotic marginal rate of substitution between consumption when young and when old or, equivalently, the gross marginal return on saving between consecutive periods. Namely, if this return is above or equal to one, the equilibrium is efficient, and if it is below one, it is inefficient.[5] Hence, in the $\omega_y \leq \omega_o$ case, where money cannot have value, there is Pareto efficiency. When money can have value ($\omega_y > \omega_o$), all equilibria (non-stationary monetary and non-monetary) are Pareto inefficient, except for the stationary monetary equilibrium: the long-run return on saving is $\omega_o/\omega_y < 1$ in the former cases and 1 in the stationary monetary equilibrium.

Is it the case that monetary equilibria, when they exist, are helpful from a welfare perspective? In particular, do they provide a Pareto improvement on the autarky allocation? Clearly, the stationary monetary equilibrium does: the initial old are happier in it, since they can sell their money for a positive amount, and all the other cohorts obtain utility equal to $2\log((\omega_y + \omega_o)/2)$, which exceeds $\log \omega_y + \log \omega_o$. Notice that the fully smoothed allocation is better than autarky for all cohorts born at $t \geq 0$, also in the case where $\omega_y < \omega_o$, but money cannot help attain this allocation. Moreover, if it could, it would make the initial old worse off.

In fact, all the non-stationary monetary equilibria do improve on autarky as well. Moreover, the monetary equilibria are ranked in the following sense: the larger is $p_{m,0} \leq \bar{p}_m$, the higher is utility for all generations. We will leave these results as an exercise for the reader; the key insight is that a higher return on saving is better for the consumer in this case.

17.2.3 Extensions: a neoclassical growth economy and policy

The basic results of the previous section extend in various ways. The essential ingredient for money to have value in this context is the lack of a (good) store of value. Alternatively, money does not have value here, if people don't need a store of value—such as when $\omega_y < \omega_o$—or a "good enough" alternative store of value is available. So, consider the overlapping-generations version of the neoclassical growth economy of Chapter 5, where people can save using physical capital. Capital would then compete with money and potentially rule out monetary equilibria.

Now, the characterization of equilibria is somewhat more cumbersome than in the endowment case. However, consider any such equilibrium, and for the preferences assume $\log c_y + \log c_o$, just as above. Recall that an equilibrium is still Pareto inefficient if, as time goes to infinity, the limit for the gross interest rate is below one. This, moreover, is an outcome that is possible in the neoclassical model. It's another good exercise to try to derive a

[5] The non-monetary equilibria described here do not have a return on saving since there is no active savings vehicle. However, we can introduce borrowing and lending, which have to equal zero in equilibrium. The gross interest rate therefore needs to adjust to ω_o/ω_y to make the young choose exactly zero borrowing.

condition under which you can see this to be true.[6] In such a case, fiat money can have value and real effects on economic activity. Capital will not be displaced entirely if production is Cobb-Douglas, so money and capital will both be used and give the same gross return of 1 asymptotically.

It is straightforward to introduce money printing into the model, e.g., as lump-sum transfers. If the government increases the money supply at a gross rate of $1+\mu$ every period, then the same type of equilibrium characterization as above obtains, with the difference that the stationary monetary equilibrium is now replaced by one where the real return on money is below one—it is $1/(1+\mu)$—and where the total value of the money stock is lower. Money printing can also be used as a source of income for the government (seigniorage) to pay for its expenditures or debts. One can thus also introduce government bonds easily. If they are in positive supply, bonds would compete with money as a store of value. Open market operations can be studied, whereby the government would change its outstanding stock of bonds over time, which in general will also affect allocations and the value of money. However, the effects on allocations only materialize if the present-value budget of a cohort is changed; otherwise, a Ricardian-equivalence theorem applies, as studied in Chapter 15. We will return to some of these issues in the next section.

Thus we can construct versions of the overlapping-generations model for which valued money co-exists with alternative means of storage if it pays the same rate of return as these alternative assets. However, the overlapping generations model still has trouble explaining why people in actual economies hold money despite there being other assets that dominate it in return.

17.3 Money in dynastic models

Dynastic models are different from overlapping-generations models in that their competitive equilibria, absent other frictions, are Pareto optimal. Thus, at least in this sense, money is not needed (as a store of value or otherwise). We will thus begin to demonstrate that indeed, money cannot have value in the basic dynastic settings. Then we introduce, one by one, additional assumptions that will give money value.

17.3.1 Fiat money has no value

We now show that in a standard infinite-horizon environment with dynastic households fiat money has no value if it only serves as a store of value. For this and the following sections, we will consider variations of a representative agent production economy with labor or one with endowments only (both covered in Chapter 5). Unlike in the section on overlapping generations, we will now explicitly consider a more general structure with government bonds as well as changes in the stock of money.

The household's preferences over consumption, c, and leisure, l, and the production of

[6] To find at least one example, look at a case where there is a closed-form solution, such as when $\omega_o = 0$ and δ, the rate of depreciation of physical capital, is 1.

consumption through labor are described by

$$\sum_{t=0}^{\infty} \beta^t u\left(c_t, l_t\right) \tag{17.6}$$

$$c_t = z\left(1 - l_t\right). \tag{17.7}$$

We have assumed a fixed time endowment of one.[7] The planning solution would hence maximize the stated utility subject to this constraint; the solution is time-independent and characterized by $zu_c(c,l) = u_l(c,l)$.

The household's budget constraint, in real terms, is

$$c_t + q_t a_{t+1} + p_{m,t} M_{t+1} + p_{m,t} B_{t+1} = w_t\left(1 - l_t\right) + a_t + p_{m,t} M_t + p_{m,t}\left(1 + i_{t-1}\right) B_t - \tau_t, \tag{17.8}$$

where a is a real asset with discount rate $q_t = 1/(1+r_t)$, and r_t is the net real return between t and $t+1$. Real wages are w, and the government imposes a lump sum tax, τ. M and B are nominal holdings of money and one-period bonds, respectively, denominated in units of money, and p_m is the price of money. Hence, an equilibrium where money has no value would have $p_{m,t} = 0$ for all t. The nominal net interest rate on bonds between t and $t+1$ is i_t.

The household is allowed to borrow or lend in real assets and nominal bonds, but cannot borrow by issuing money,

$$M_t \geq 0. \tag{17.9}$$

The government's budget constraint (where we abstract away from real expenditures) is given by

$$p_{m,t}(M_{t+1} - M_t) + p_{m,t} B_{t+1} + \tau_t = p_{m,t}\left(1 + i_{t-1}\right) B_t. \tag{17.10}$$

In equilibrium, the representative household holds outstanding fiat money and nominal government debt and real assets are zero, $a_t = 0$. Thus, the resource constraint is satisfied: $c_t = z(1 - l_t)$.

Let $m_t \equiv p_{m,t} M_t$ and $b_t \equiv p_{m,t} B_t$ denote the real value of money and bonds and define $\hat{a}_t = a_t + m_t + b_t(1 + i_{t-1})$ as total real wealth at the beginning of the period. By using this definition, along with some algebra, we can rewrite the household's budget as

$$c_t + q_t \hat{a}_{t+1} + \left[\frac{p_{m,t}}{p_{m,t+1}} - q_t\right] m_{t+1} + \left[\frac{p_{m,t}}{p_{m,t+1}} - q_t\left(1 + i_t\right)\right] b_{t+1} = w_t\left(1 - l_t\right) - \tau_t + \hat{a}_t. \tag{17.11}$$

In this budget constraint, $\hat{a}$ (and no other variable) is used to save from t to $t+1$. The terms involving m_{t+1} and b_{t+1} are "static": they are losses (gains) at t to the extent the expressions in square brackets are positive (negative). Thus, first, the household would not want to hold money if its real return, $p_{m,t+1}/p_{m,t}$, is below the real interest rate $1/q_t$. Second, the household would be able to attain unbounded consumption for given wealth and prices if the real return on bonds, $(1 + i_t)p_{m,t+1}/p_{m,t}$ were not equal to the real interest rate: if the return is higher (lower), the consumer could obtain unbounded resources by raising

[7] Note that in the other chapters of this textbook, we use ℓ for labor supply. In the context of this chapter, we have $l + \ell = 1$.

(lowering) b_{t+1} without bound. Thus, for an equilibrium with positive money and finite bond holdings to exist, the following needs to hold:

$$q_t = \frac{p_{m,t}}{p_{m,t+1}} = \frac{p_{m,t}}{p_{m,t+1}} \frac{1}{1+i_t}. \tag{17.12}$$

These no-arbitrage conditions mean that we can write the budget constraint in more compact form as

$$c_t + q_t \hat{a}_{t+1} = w_t (1 - l_t) - \tau_t + \hat{a}_t. \tag{17.13}$$

We see immediately from equation (17.12) that money cannot be held in positive amounts—it cannot have real value, i.e., p_m will have to equal 0—if the nominal interest rate on bonds is positive. If $i_t > 0$, money would then be dominated in return by bonds and be a pure loss to hold in positive amounts. Second, if i_t is zero at all times (or nominal bonds are not available in the economy), then money (or bonds, which are now equivalent to money) cannot have value either.[8] To see this, note that the consumer's optimization problem collapses to a problem with one asset with real return $1/q_t$, equation (17.13). From Chapter 4 we know that for this problem the transversality condition is a necessary condition for optimality. Applying the condition to the real value of money holdings we obtain

$$0 \geq \lim_{T\to\infty} q_0 q_1 \dots q_T m_T = \lim_{T\to\infty} \frac{p_{m,0}}{p_{m,1}} \frac{p_{m,1}}{p_{m,2}} \dots \frac{p_{m,T-1}}{p_{m,T}} p_{m,T} M_T = p_{m,0} \lim_{T\to\infty} M_T. \tag{17.14}$$

Thus, if fiat money is not vanishing (i.e., being withdrawn) in this economy, money cannot have value earlier on. In other words, for money to have value in this economy it must disappear in the limit. We will return to this point.

So far we have been interested in whether fiat money can have value, that is, whether the price of money in terms of goods can be positive. In the following we will mainly deal with environments for which money does have value, that is, $p_m > 0$. For these environments we are often interested in the price of goods in terms of money, that is, the price level, $P = 1/p_m$, and the rate at which the price level is changing, that is, the gross inflation rate, $1 + \pi = P'/P = p_m/p'_m$. For nominal bonds, this means that we will frequently write the expression for their return as the *Fisher equation*:

$$1 + i_t = (1 + r_t)(1 + \pi_t), \tag{17.15}$$

which states that the gross nominal interest rate equals the gross real interest rate times the gross inflation rate. In our perfect foresight environment, the Fisher equation can be seen as an arbitrage condition involving real and nominal bonds.

17.3.2 Fiat money with reduced-form liquidity services has value

There are different ways to give value to money by assuming, in a reduced-form way, that it matters to consumers. We now briefly look at them, one by one. Because government bonds, from the perspective of the consumer, are identical to borrowing and lending—both b and a can be held in positive as well as negative amounts and hence will yield the same return—we will only keep a. We will sometimes refer to the nominal interest i_t, which as before means the money return at $t + 1$ on a nominal bond bought at t.

[8] Recall that nominal interest rates at zero were observed for an extended period in the aftermath of the Great Recession, so this case is not just a theoretical one.

The cash-in-advance model

We first impose the constraint that money has to be used in transactions; in particular, we assume that goods can only be purchased using money. This constraint, commonly known as a cash-in-advance (CIA) constraint, from Clower (1967), gives rise to an equilibrium where money has value and is dominated in return by other assets.

Let us first incorporate the CIA constraint into the household's budget constraint:

$$c_t \leq p_{m,t} M_t \tag{17.16}$$

$$q_t a_{t+1} + p_{m,t} M_{t+1} = a_t + p_{m,t} M_t - c_t + w_t (1 - l_t) - \tau_t. \tag{17.17}$$

The first expression is the CIA constraint and states that money is needed to buy goods. The second expression states that money that is not spent today can be saved, holding money or the asset a. The two expressions combined also state that current earnings are not available contemporaneously for consumption.

Consider now the modified consumption-savings problem of the representative agent

$$\max_{\{c_t, l_t, a_{t+1}, m_{t+1}\}} \sum_{t=0}^{\infty} \beta^t u (c_t, l_t)$$

subject to

$$c_t \leq m_t \tag{17.18}$$

and

$$c_t + q_t a_{t+1} + \frac{p_{m,t}}{p_{m,t+1}} m_{t+1} = a_t + m_t + w_t (1 - l_t) - \tau_t, \tag{17.19}$$

where we have again replaced nominal money balances with real balances. The first-order conditions for the household's problem are

$$u_1 (c_t, l_t) = \lambda_t + \mu_t, \tag{17.20}$$

$$u_2 (c_t, l_t) = \lambda_t w_t, \tag{17.21}$$

$$\lambda_t q_t = \beta \lambda_{t+1}, \tag{17.22}$$

and

$$\lambda_t \frac{p_{m,t}}{p_{m,t+1}} = \beta (\lambda_{t+1} + \mu_{t+1}), \tag{17.23}$$

where $\beta^t \mu_t$ and $\beta^t \lambda_t$ are the Lagrange multipliers on the CIA constraint (17.18) and budget constraint (17.19), respectively, and $u_i(\cdot, \cdot)$ is the partial derivative with respect to the ith argument.

Now consider a stationary equilibrium for a constant money stock M and labor productivity, z. In the equilibrium the household holds all of the money, net assets are zero, $a = 0$, consumption and leisure satisfy $c = z(1 - l)$, and the real wage equals the marginal product of labor, $w = z$. From the FOCs for consumption and leisure, it follows that the Lagrange multipliers are constant, and from the FOC for assets, it follows that the discount rate on assets is equal to the discount factor, $q = \beta$.

If the CIA constraint is binding with a constant c and M, then the price of money is constant at $p_m = c/M$. As there is no inflation, the nominal interest rate satisfies $q(1+i) = 1$. From the FOC for real balances, it then follows that the Lagrange multiplier on the CIA is positive, $\mu > 0$. Combining the FOC for consumption and leisure, we obtain

$$\frac{u_2(c, 1-c/z)}{u_1(c, 1-c/z)} = z\frac{\lambda}{\lambda+\mu} = \beta z, \tag{17.24}$$

where we have also used the Euler equation for real balances, (17.23). Equation (17.24) determines the steady-state consumption level from which all remaining equilibrium variables can be obtained.

Notice that the CIA must be binding in the stationary equilibrium. To see why, suppose it is not binding, that is, $c < p_{m,t}M$ and $\mu = 0$. From the first-order condition for real balances, it then follows that the return on money is equal to the interest rate. But this implies that the price of money is increasing over time. So the CIA constraint remains satisfied, but the TVC is now violated, as shown in the previous section.

Also, notice that compared to the same economy without the CIA constraint, the marginal rate of substitution between leisure and consumption is less than the marginal rate of transformation: $\beta z < z$. But this means that in the CIA economy leisure is higher and consumption is lower. The real wage effectively received by the household is lower in the CIA economy because today's labor income can only be used tomorrow, and the return on saving the labor income is less than the real interest rate. Is there a way to fix the problem?

Suppose the government imposes a nominal lump sum tax on the representative household, T_t, to be paid using money. Also, assume that money is withdrawn at a constant rate, $M_{t+1}/M_t = \gamma < 1$. Then the implied real lump sum tax is $\tau_t = p_{m,t}(M_t - M_{t+1}) = (1-\gamma)p_{m,t}M_t$. We can still find a stationary equilibrium with a binding CIA constraint where the price of money increases as the money stock shrinks and the value of the total money stock is constant, $M_t p_{m,t} = c$. If the nominal stock changes at a gross rate γ, then the value of each unit of money changes at $1/\gamma$, and the discount rate for real balances is γ; for $\gamma > 1$, this means a constant rate of inflation and for $\gamma < 1$ a constant rate of deflation.

We can, in particular, choose $\gamma = \beta$ such that the return on assets and money is equalized. For this policy the TVC is satisfied since the nominal money stock is vanishing in the limit. This result—that withdrawing money from the economy at the rate of discount makes the equilibrium optimal, and constitutes optimal monetary policy—is known as the *Friedman rule*. Milton Friedman cast this policy in terms of "paying interest on money", which is something a central bank could engineer. With the money stock shrinking at rate β, people are not constrained in the use of money, and similarly if money paid interest and were therefore identical to bonds, they would not be constrained either.

The CIA model of money does not satisfy Wallace's dictum: though we have not demonstrated it, money will always be valued here because it has been hard-wired to be equal to consumption (in real terms, and hence its value could not be zero, or else consumption would be zero). Money is a store of value in the CIA model, but one that is worse than bonds and other assets: its value derives from it being required to buy goods. Money is thus used as a medium of exchange, though the exchange is not explicitly modeled. One way to describe the CIA constraint is that it imposes a *quantity equation* of sorts: the equation $VM = Py = Pc$ is met by assumption with a velocity of 1, $V = 1$.

The quantity theory in the data

The defining characteristics of the Quantity Theory of Money (QTM) are the presence of a stable demand for money and that money growth and nominal GDP growth move one for one. Suppose that money is neutral over the long run, that is, real activity is independent of money growth over the long run. Then price inflation will move one-for-one with money growth in excess of real GDP growth over the long run if the demand for money is stable and proportional to the nominal transaction volume.

For an empirical evaluation of QTM one thus needs measures of the money stock, the transaction volume, quantities and prices, and the opportunity cost of holding money. From the point of view of money as a means to execute transactions, one can think of various measures. Standard measures of money published by central banks include M1, consisting of currency and checkable demand deposits, and M2, which adds savings deposits to M1. Regulatory changes and technical advances may affect what should be included in a measure of money, see Lucas and Nicolini (2015). For example, before the 1980s, Regulation Q in the U.S. imposed limits on the ability of banks to pay interest on accounts. Once Regulation Q was abolished, banks started to offer new interest-bearing liquid accounts, e.g., money market deposit accounts with limited transaction features. In the mid-1990s, IT improvements made moving funds between demand deposits and money market accounts easier (SWEEP accounts), making the latter close payment substitutes. Payments from bank accounts have also been made easier using electronic transfers that can be initiated using mobile phones. Bitcoin and other digital currencies may represent additional future means of payment. Thus what should be included in a measure of money changes over time, which affects the stability of money demand as defined by a fixed measure of money. Furthermore, the M1 and M2 measures listed are simple sums of their various components, but these components may well differ in their ability to perform transactions. Divisia indices have been proposed for constructing aggregate money stocks; see Barnett (1980). This approach is analogous to how we aggregate different final goods into an aggregate output measure like GDP. The standard measure of transactions in the literature is GDP, but obviously, many transactions precede the purchases of final goods in the economy: there are usually multiple stages of production. Thus, using GDP as a sufficient statistic for the transaction volume implicitly assumes that the production structure is not changing much over time. Finally, various short-term interest rates have been used for the opportunity cost of holding money. For the U.S. that includes the Federal Funds rate, the rates on short-term commercial paper or short-term U.S. Treasuries. In the following, we study how well the QTM holds up for the U.S. for the period 1901 to 2023 using M2 as a measure of the money stock, GDP as a measure of transactions, and the 6-month commercial paper rate as a measure of the opportunity cost of holding money. Using other measures of money and interest rates produces very similar results.

We first consider the long-run relation between M2 growth and inflation, following Sargent and Surico (2011). For this purpose, we calculate long-run movements of M2 growth and inflation as 15-year symmetric moving averages. In Figure 17.3 we plot the filtered M2 growth rates and inflation for annual data from 1902–2023 for five sub-

samples: 1902–1928, 1929–1954, 1955–1983, 1984–2005, 2006–2023. The first period precedes the Great Depression, the second period covers the Great Depression and World War II, the third period covers the post-WW-II period including the inflationary 1970s, the fourth period covers what has been called the Great Moderation and the adoption of inflation targeting among advanced economies, and the fifth period covers the Great Recession and the policy of Quantitative Easing (QE). We see that inflation moves roughly one-for-one with money growth, but that there is notable variation in that relation across subsamples. On the one hand, during the period covering high inflation in the late 1960s and 1970s, inflation appears to respond strongly to changes in money growth. On the other hand, during the Great Moderation and inflation targeting, the response of inflation to changes in money growth is much weaker, and during the QE period, inflation appears to decline as money growth increases.

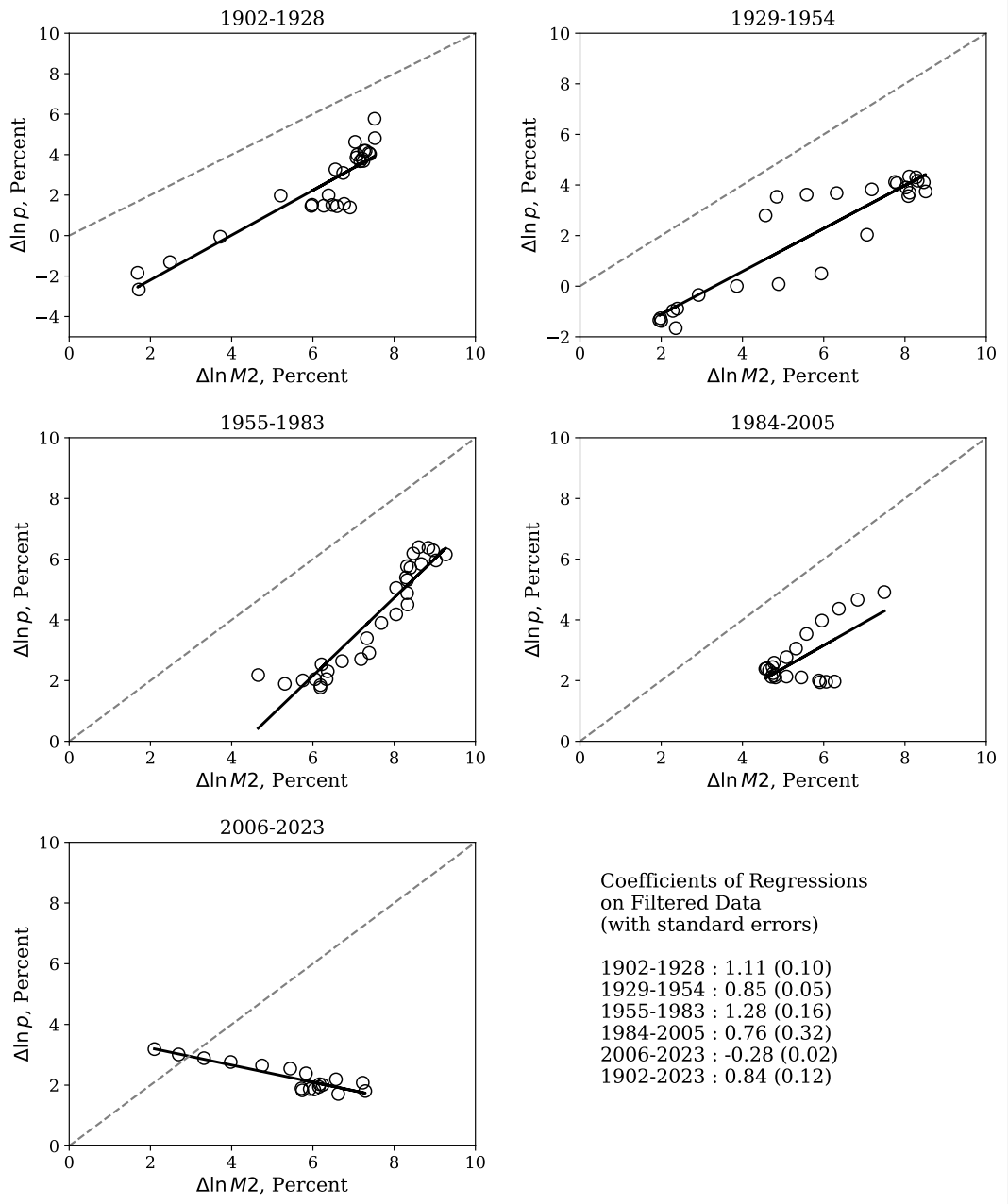


Figure 17.3: Inflation and money growth

Notes: Each panel plots the filtered inflation rates against the filtered M2 growth rates (circles), and contains the 45-degree line (dashed line) and the fitted values from an OLS regression of inflation on money growth (thick line). The estimated OLS coefficients for the sub-samples are in the lower right-hand corner, with heteroskedasticity and auto-correlation corrected standard errors in parentheses.

We now consider the long-run stability of money demand. In Figure 17.4 we plot the M2-GDP ratio and the 6-month commercial paper rate for both the actual data and their 15-year symmetric moving averages. At first inspection, Figure 17.4 seems

to provide evidence for the long-run stability of money demand despite the large and persistent deviations of the variables from their long-run trends. Whenever the filtered short rate increases, the filtered M2-GDP ratio declines. This stable relationship for the filtered data disappears, however, when we look at the sub-samples as in Figure 17.3. Now, the OLS regression coefficients of the M2-GDP ratio on the short rate can be either positive or negative, and they are usually not significant. A cautious summary is that the evidence on QTM is mixed.

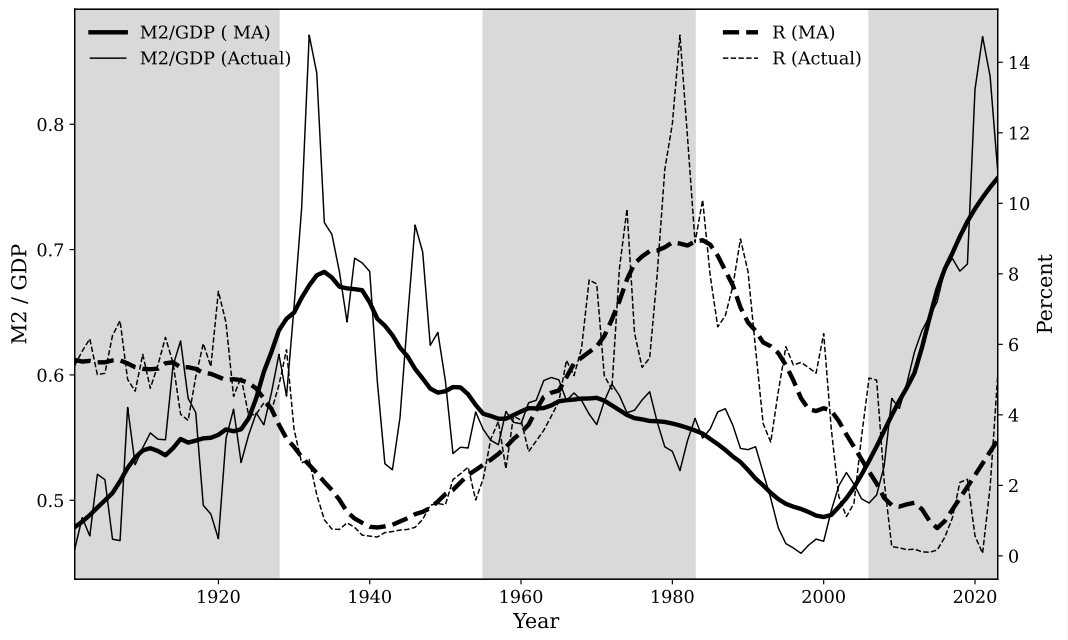


Figure 17.4: Money demand

Notes: The thin lines are the annual data for the ratio of M2 to nominal GDP (solid, left axis) and the 6-month Commercial Paper (CP) rate (dashed, right axis). The thick black lines are the 15-year symmetric moving averages of the M2-GDP ratio (solid) and the CP rate (dashed). Shading and its absence sets apart the five subsamples from Figure 17.3.

Besides a CIA constraint, there are other ways to introduce money into a standard representative agent frictionless economy. Next, we consider two prominent examples. The first example posits that transactions are costly, but less so if one holds money. The second example represents any inherent value to money by incorporating it directly into the utility function.

Money reduces transactions costs

Consider the budget constraint (17.19), but assume that the purchase of consumption goods involves a real cost in terms of the consumption good,

$$c_t + q_t a_{t+1} + \frac{p_{m,t}}{p_{m,t+1}} m_{t+1} = a_t + m_t + w_t (1 - l_t) - \tau_t - \psi \left(\frac{c_t}{M_{t+1}/P_t} \right) c_t, \tag{17.25}$$

where $\psi(x) = \kappa x^{\eta}$ is an increasing function, $\kappa, \eta > 0$. Thus transaction costs are increasing in consumption, decreasing in current real balances, which we recall are $M_{t+1}/P_t = m_{t+1}p_{m,t}/p_{m,t+1}$, and homogeneous of degree one in consumption and real balances jointly. Here, real balances chosen in the current period facilitate transactions, unlike in the CIA model where real balances carried over from the previous period are required for transactions.

Using a subscript t on ψ and its derivative to indicate evaluation at $c_t/\left(m_{t+1}p_{m,t}/p_{m,t+1}\right)$, we obtain first-order conditions for the household's problem as follows:

$$u_1\left(c_t, l_t\right) = \lambda_t \left[1 + \psi_t' \cdot \frac{c_t}{m_{t+1}p_{m,t}/p_{m,t+1}} + \psi_t\right],$$

$$u_2\left(c_t, l_t\right) = \lambda_t w_t,$$

$$\lambda_t q_t = \beta \lambda_{t+1}, \tag{17.26}$$

and

$$\lambda_t \frac{p_{m,t}}{p_{m,t+1}} \left[1 - \psi_t' \cdot \left(\frac{c_t}{m_{t+1}p_{m,t}/p_{m,t+1}}\right)^2\right] = \beta \lambda_{t+1}, \tag{17.27}$$

where $\beta^t \lambda_t$ is the Lagrange multiplier on the budget constraint (17.25). Combining the FOCs for the real asset and real balances, (17.26) and (17.27), we obtain

$$1 - \frac{q_t}{p_{m,t}/p_{m,t+1}} = \psi_t' \cdot \left(\frac{c_t}{m_{t+1}p_{m,t}/p_{m,t+1}}\right)^2. \tag{17.28}$$

Using the Fisher equation (17.15) for the nominal interest rate on the left-hand side and using the functional form for ψ on the right-hand side we arrive at an expression,

$$M_{t+1} = P_t c_t \left[\frac{1}{\kappa\eta} \frac{i_t}{1 + i_t}\right]^{-1/(1+\eta)}, \tag{17.29}$$

which relates the demand for money to the value of transactions and the opportunity cost of holding money, the nominal interest rate. This expression is in line with the usual expressions for money demand, unlike in the simple CIA model above, for which money demand is interest-inelastic.[9] That is, real money demand is equal to consumption (or output, if output is defined net of transactions costs) times the inverse of velocity, an expression that is decreasing in the nominal interest rate. We have so far described this model only from a single consumer's perspective. To solve for a general equilibrium, we would need to specify assumptions on production, taxes, and market clearing.

Money in the utility function

We have just assumed that real balances are for some reason helpful in executing consumption purchases. We could take an even more reduced form approach and assume that real balances are valuable, period. For this purpose, we simply include real balances in the utility function

[9] More elaborate versions of the CIA model also feature interest-elastic money demand.

of the representative agent, rather than modifying the budget constraint. So, assume that preferences are

$$\sum_{t=0}^{\infty} \beta^t \left[u\left(c_t, m_{t+1} \frac{p_{m,t}}{p_{m,t+1}} \right) + v\left(l_t \right) \right], \tag{17.30}$$

where the second argument of u is simply the real amount of money purchased at t, M_{t+1}/P_t, but rewritten in terms of the real balanced in the beginning of next period and the prices of money in the two consecutive periods. The specific functional form used is convenient but of course not the only possibility. The FOCs for the household's consumption-savings problem with preferences (17.30) and budget constraint (17.19) are

$$u_{1,t} = \lambda_t, \tag{17.31}$$

$$v'_t = \lambda_t w_t, \tag{17.32}$$

$$\lambda_t q_t = \beta \lambda_{t+1}, \tag{17.33}$$

and

$$\lambda_t \frac{p_{m,t}}{p_{m,t+1}} = u_{2,t} \frac{p_{m,t}}{p_{m,t+1}} + \beta \lambda_{t+1}, \tag{17.34}$$

where we have suppressed the arguments of the derivatives of u and v for compactness ($u_{i,t}$ again denotes the partial derivative with respect to ith argument), and subscript t denotes time-dependence of the arguments. Another equation is also relevant: our budget constraint (17.19) does not explicitly feature nominal government bonds paying interest i_t between t and $t+1$, but that is because of the no-arbitrage condition between real and nominal bonds that yields the Fisher equation, $q_t(1+i_t) = p_{m,t}/p_{m,t+1}$. Using this equation together with equations (17.33) and (17.34), we obtain $1 = u_{2,t}/u_{1,t} + 1/(1+i_t)$. This delivers an expression that is very similar to equation (17.29):

$$\frac{u_{2,t}}{u_{1,t}} = \frac{i_t}{1+i_t}. \tag{17.35}$$

If we assume that $u(c,m)$ is a homothetic function, that is, it is a monotone transformation of a homogeneous of degree one function, we again obtain a static money demand equation that relates the ratio of real balances to consumption, m/c, to the opportunity cost of holding real balances, $i/(1+i)$, like (17.29). As in the transactions-costs case, the stationary general equilibrium allocation is then obtained by adding assumptions on production, taxes, and market clearing.

17.3.3 Policy and the value of money in the reduced-form models

We will now illustrate some key features of the reduced-form models of money.

Stationary equilibria: quantity theory and optimality

All of our reduced-form models of fiat money give rise to a well-defined demand for real balances. In the CIA model, this demand for real balances is independent of the nominal interest rate, equation (17.18) with equality. But for the models where real balances reduce

transaction costs or provide utility directly, this demand may depend on the nominal interest rate, e.g., equations (17.29) and (17.35). Since the environment in our examples is stationary so far we have studied their stationary equilibria for which quantities and prices are constant. In these equilibria real balances are constant and therefore the price of money is inversely proportional to the money stock. This reflects the quantity theory which is usually taken to state that the price level is proportional to the money stock.

The Friedman rule In the CIA model, we have shown that in general the stationary equilibrium outcome is sub-optimal relative to the economy without frictions, that is, without the CIA constraint. We also saw that the outcome of the frictionless economy can be recovered if the money stock is shrinking at a rate such that the price of money is increasing and the real rate of return on money is equal to the real discount rate. In other words, the price level is falling, and there is deflation at the real interest rate. Furthermore, from the Fisher equation, it then follows that the nominal interest rate is zero. This is the Friedman Rule: if money is dominated in return but fills a function in the economy, then optimal policy is to reduce the opportunity cost of holding money to reduce distortions. For the CIA model, such a policy can eliminate the distortion, but for the two other reduced-form models of money, the full optimum can only be reached exactly if additional "satiation assumptions" are made. This means that for some finite value of real balances, transaction costs are zero in the first economy, and preferences have a bliss point in real balances in the second.

Equilibrium indeterminacy

A feature of monetary economies is that they naturally give rise to multiple equilibria. This is mainly because forsaking real resources and buying money today is not advantageous if it loses value entirely in the next period (i.e. if others are not willing to buy it): the price of money today depends naturally on the price of money in the future. In the overlapping-generations economy of section 17.2, there is always an equilibrium where money never has value because money is used only as a savings vehicle. As we shall see here, reduced-form models often also deliver indeterminacy.

Hyperinflation In the reduced-form models, equilibria where money never has value can occur but only under special assumptions on preferences/transactions costs.[10] There may, however, exist non-stationary monetary equilibria. Consider, for example, the CIA model from section 17.3.2. Combine equations (17.20)-(17.23), assuming that the CIA constraint is binding, and normalize $z = 1$. We obtain

$$u_2\left(m_t, 1 - m_t\right) \frac{p_{m,t}}{p_{m,t+1}} = \beta u_1\left(m_{t+1}, 1 - m_{t+1}\right). \tag{17.36}$$

Now assume that through lump-sum transfers (negative taxes) the money stock changes at the constant gross rate $\gamma > 1$, $M_{t+1} = \gamma M_t$, and we arrive at a first-order difference equation

[10] In the CIA model, a non-monetary equilibrium would mean zero consumption, which is only possible if utility is bounded below in consumption. In the model where money reduces transaction costs, these costs would need to be bounded at zero real money balances, and in the model where money enters the utility function, this function must be bounded below in real balances.

in real balances, m_t,

$$\gamma u_2 (m_t, 1 - m_t) \frac{m_t}{m_{t+1}} = \beta u_1 (m_{t+1}, 1 - m_{t+1}), \tag{17.37}$$

with no initial condition for real balances. This makes clear how expectations fundamentally drive equilibrium outcomes. One solution is a steady state, but this may not be the only possibility. To make this point particularly sharp, suppose that leisure and consumption are separable and that the marginal utility of leisure is constant and equal to 1: with a slight abuse of notation, $u(c, l) = u(c) + l$. This yields

$$\gamma m_t = \beta m_{t+1} u' (m_{t+1}). \tag{17.38}$$

If in addition $\lim_{m \to 0} mu'(m) = 0$, then this expression defines two steady states: one with positive real balances, $\gamma/\beta = u'(\bar{m})$, and a limiting steady state where money has no value and real balances are zero. Furthermore, for any initial real balances $0 < m_0 < \bar{m}$ the path of real balances defined by equation (17.38) converges to the steady state with zero real balances.[11] Given M_0 an initial m_0 corresponds to a price $p_{m,0}$, and $p_{m,0} < \bar{p}_{m,0}$ for $m_0 < \bar{m}_0$. Thus, for any initial $p_{m,0} < \bar{p}_{m,0}$ the price of money converges to zero faster than the money stock increases, and the value of real balances vanishes in the limit.[12]

Local indeterminacy In a different environment, the same policy of constant money growth through lump-sum transfers can result in a continuum of nonstationary equilibria that all converge to the unique stationary equilibrium. For a version of the reduced-form model with money in the utility function from section 17.3.2, Obstfeld (1984) shows that the local dynamics of perfect foresight paths at the steady state allow for a continuum of paths that all converge to the steady state if consumption is a normal good and the magnitude of the elasticity of marginal utility of consumption with respect to real balances is sufficiently small. For all of these equilibria, the price level relative to the money stock remains bounded. Below we will study the interaction of monetary and fiscal policy and explore how policy modifications can lead to a unique equilibrium.

Different monetary rules

The above discussion contains an implicit assumption about the conduct of monetary policy: the monetary authority selects a path for the money supply, a sequence that will be given to the economy, and for which one can then examine the set of implied equilibrium allocations. We learned—for the economies we looked at—that for a money supply path featuring growth at a constant rate, there is a unique steady-state equilibrium in which the rate of inflation equals the inverse of the money growth rate. We also learned that other equilibria may

[11] Linearizing the dynamic system around $m = \bar{m}$, we see that the system is locally unstable. Linearizing around $m = 0$, however, delivers stability, provided marginal utility is sufficiently large, $u'(0) > \gamma/\beta$. Note also that along paths where $m_t \to 0$, the real return on money is always less than in the stationary equilibrium with positive real balances, and thus the CIA is binding.

[12] The example in this section relies on specific assumptions on utility. Under other assumptions, there is a unique equilibrium. If $u(c) = \log c$, which violates the condition stated as $\lim_{m \to 0} mu'(m) = 1 > 0$, equation (17.38) allows us to solve uniquely for m_0, m_1, and so on.

exist under some conditions. For the case in which there are government bonds, the nominal interest rate consistent with the money stock sequence would follow from the Fisher equation. However, in most economies of today, monetary policy is not described this way; rather, the more appropriate description of the conduct of monetary policy is that of a "choice of a sequence of nominal interest rates", possibly associated with a rule such as the Taylor rule as described below. We now discuss how to define equilibria where the central bank directly chooses interest rates (and the money supply path becomes endogenous).

Interest rate rules

Sims (1980b) popularized the use of vector autoregressions (VARs) as a way of studying the effects of exogenous shocks on macroeconomic time series while imposing minimal assumptions to identify the exogenous shocks (see Chapter 8). Early VAR applications that studied monetary policy as a source of economic fluctuations were influenced by QTM arguments. Sims (1972), in a small-scale VAR of the U.S. economy with real GDP and nominal money finds a quantitatively large contribution of money shocks to output fluctuations. Later work considered larger-scale VARs, and once short-term interest rates were added to the list of variables, interest rate shocks replaced money shocks as a source of output fluctuations. Bernanke and Blinder (1992) then argued that monetary policy actions set a particular short-term interest rate, namely the Federal Funds rate, the overnight interest rate in the market for interbank loans, and thus interest rate shocks reflect the impact of monetary policy. Contemporaneously, Goodfriend (1991), based on a reading of U.S. monetary policy implementation at the Federal System, argues that

> "Except for the period from 1934 to the end of the 1940s when short-term interest rates were near zero or pegged, the Fed has always employed either a direct or an indirect Federal Funds rate policy instrument." (p.8)

In particular, he observes that the Federal Funds rate is adjusted infrequently and that rate changes usually occur in a sequence of consecutive small steps.

One possibility is to set interest rates according to a function of observed macroeconomic data. The Taylor (1993) rule is the canonical example of such an interest rate rule. An interest rate rule of one kind or another is now an integral part of most quantitative monetary models. Nevertheless, practitioners of monetary policy, for example, Bernanke (2015), frequently argue that the guidance provided by interest rate rules is limited due to, among others, the difficulties in assessing the theoretically appropriate conditioning variables, such as output gaps and natural real rates, in real time.

Price level indeterminacy under pure interest rate rules First, consider a monetary policy that would just specify a sequence of nominal interest rates. For a long time, such a policy has been viewed as perilous, as it leaves the price level indeterminate. To see why, consider a pure interest-rate peg in the context of the general-equilibrium cash-in-advance model from Section 17.3.2. In particular, assume a path for the nominal interest rate for which $i_t > 0$. As in our discussion of hyperinflation, we specialize utility to be linear in

leisure; in particular, let period utility be $u(c) + l$ and let labor productivity z be equal to one. Then recall that the equilibrium conditions, which we only studied in a steady-state version in section 17.3.2, will read (in the order stated in that section) $c_t \leq p_{m,t}M_t$, $u'(c_t) = \lambda_t + \mu_t$, $1 = \lambda_t$, $q_t\lambda_t = \beta\lambda_{t+1}$, and $\lambda_t p_{m,t}/p_{m,t+1} = \beta(\lambda_{t+1} + \mu_{t+1})$. From this follows that $q_t = \beta$ and that $p_{m,t}/p_{m,t+1} = \beta(1 + i_t)$, using the Fisher equation, for all $t \geq 0$. Thus, the interest-rate peg determines inflation. This also means, using the last equilibrium condition, that $p_{m,t}/p_{m,t+1} = \beta u'(c_{t+1})$, thus, pinning down c_{t+1} for all $t \geq 0$. Thus, conditional on the initial price of money, $p_{m,0}$, the peg delivers a unique deterministic equilibrium from the second period on.

However, $p_{m,0}$ is not determined. Suppose that the CIA constraint in the initial period does not bind, that is, the initial CIA multiplier is zero, $\mu_0 = 0$, and consumption is determined by $u'(c_0^*) = 1$. For the given initial money stock M_0 any positive initial price of money $p_{m,0} > p_{m,0}^* = c_0^*/M_0$ will then satisfy the CIA constraint and represent an equilibrium. Alternatively, if $p_{m,0} \leq p_{m,0}^*$ then the initial CIA constraint binds, and the initial price together with the CIA constraint determines initial consumption, c_0, and the FOC for consumption determines the CIA multiplier, μ_0. So, any initial positive price of money indexes an equilibrium, and there is real, and not just nominal, indeterminacy.

Interest rate rules with a target Now consider amending monetary policy with an *interest rate rule*: a function relating the set interest rate to the price level (or, more commonly in practice, to the inflation rate, as in the Taylor rule)

$$i_t = \phi\left(\frac{P_t}{\widetilde{P}_t}\right). \tag{17.39}$$

Here, $\widetilde{P}_t$ is a target price level, which can change over time. The idea is that, under appropriate restrictions on the function ϕ, the basic price level indeterminacy disappears.[13]

How does the addition of the interest rate rule affect the determination of the price level? Let us go through the case of money in the utility function. Suppose, for a moment, that the real side of the economy is not affected by nominal variables. That would mean that the path for r_t can be taken as given. To make matters even simpler, suppose that r_t is constant: $r_t = r$ for all t. The Fisher equation then reads $1 + i_t = (1 + r)\,(P_{t+1}/P_t)$. It follows that if the policy attains its target price level in every period, the target interest rate is $\tilde{i} = \phi(1)$, and the target price level must change at a constant target inflation rate, $(1 + \tilde{i})/(1 + r) = \widetilde{P}'/\widetilde{P} = 1 + \widetilde{\pi}$.

Combining the Fisher equation with the interest rate rule we can then write the evolution of the price level deviation from target as

$$\frac{P_{t+1}}{\widetilde{P}_{t+1}} = \frac{P_t(1 + i_t)/(1 + r)}{(1 + \widetilde{\pi})\widetilde{P}_t} = \frac{1 + \phi\left(P_t/\widetilde{P}_t\right)}{1 + \widetilde{i}}\frac{P_t}{\widetilde{P}_t}, \tag{17.40}$$

Equation (17.40) thus defines a first-order difference equation for the nominal price level relative to its target, without an initial condition. Its local dynamics around a steady

[13]Note that from now on the discussion will be in terms of the price level, P, and no longer in terms of the price of money, p_m.

state of 1 are straightforwardly determined by the use of a first-order Taylor approximation: $P_{t+1}/\widetilde{P}_{t+1} - 1 \approx \left[1 + \phi'(1)/(1+\widetilde{i})\right]\left(P_t/\widetilde{P}_t - 1\right)$. If $\phi'(1) > 0$, the local dynamics are determinate: there is a unique bounded solution with $P_t = \widetilde{P}_t$ for all t.[14] If, on the other hand, $\phi'(1) < 0$, then there is indeterminacy: a continuum of paths are consistent with convergence to the given steady state.[15] In conclusion, we see that if the central bank uses a rule that raises the nominal interest rate in response to an increase of the price level relative to its target, then the indeterminacy of nominal prices is no longer a problem. The corresponding demand for nominal money then follows from the money demand equation (17.35).

The discussion here has taken as given an exogenous path for r_t that, moreover, is constant. The arguments extend to an exogenous path for r_t that converges to a constant. How restrictive is the assumption that the path for r_t is exogenous? We know that the Euler equation for the real rate when money is in the utility function from equations (17.31) and (17.33) will, in general, involve nominal variables and hence make the analysis more involved. A full equilibrium treatment would require specification of the production side, and the equilibrium conditions would then need to be solved for. One possibility is that we again assume that production is linear in labor, $c = z(1-l)$, and that the government does not consume. Then the consumer's first-order condition for leisure, $zu_1(c_t, M_{t+1}/P_t) = v'(l_t)$, can be used directly to solve for time-independent $c_t = c$ and $l_t = l$ if we also assume that u is separable in consumption and real money balances. Hence, we can obtain $r_t = r$.[16] For the more general case of non-separability, one would need to examine the joint system of prices and real money holdings. It is of course possible to do so and to establish joint conditions on ϕ and u such that the equilibrium is determinate. The details are not important, so we omit them here. Nevertheless, dealing with the general case is relevant since a utility function where money enters separately from consumption is hard to motivate: after all, money's role is meant to be tied to the purchase of consumption goods.

Paying interest on money Recall Friedman's proposal: to pay interest on money. It is indeed conceivable for a central bank to pay interest on money, and it is even a policy in practical use, because many central banks pay interest on reserves, which are commercial bank accounts with a central bank and are part of what is considered money. Let i_m denote the interest paid on money. If $i_m = i$ at all times, money is equivalent to bonds in financial terms for the household and no reduced-form approach is necessary for obtaining a value of money. If $i_m < i$, money is dominated in return by bonds and a reduced-form demand would need to be added for money to have value in equilibrium. Now, we can add interest on money in the budget constraint (17.11) with total real wealth $\hat{a}$ and the static opportunity costs of holding bonds or money, and the real opportunity cost of holding each unit of money becomes $(i - i_m)/(1+i)$ multiplying M'/P. Using an $i_m \in (0, i)$ as an additional policy variable would not add conceptually to the rest of the analysis in this chapter, which is why

[14] Here we focus on determinacy of bounded paths for the economy. Unbounded equilibria would need to be ruled out by other arguments.

[15] The knife-edge case $\phi'(1) = 0$ leads to indeterminacy as well: a continuum of non-diverging price ratios $P_t/\widetilde{P}_t$ are possible (constant) solutions to the equilibrium equations.

[16] This argument also works if z depends on time but is converging to a constant, in which case r_t converges to a constant.

our benchmark maintains $i_m = 0$.

The cashless limit

An important part of the New Keynesian model, to which Chapter 18 is devoted, deals with price level determination in the absence of a demand for money: money balances are typically omitted from that model. The price level is still nominal—set in units of currency—and serves a different role: it is the unit in which prices are assumed to be sticky. Price stickiness is not the subject of the present chapter, however; suffice it to say that in the New Keynesian model, a firm's price *in dollar terms* cannot be changed freely. But how is it possible to have nominal variables in the model without having money? We now discuss how Woodford (2003a) motivates this approach.

So consider again the model with money in the utility function. We just saw that a framework in which u is additively separable in consumption and real money balances is consistent with price-level determinacy under an interest rate rule, which is also the standard assumption about monetary policy made in the New Keynesian model. Moreover, with separable preferences, money balances do not matter for the determination of either real variables or the price level, but they are simply given *residually* from the money demand relationship (17.35). This economy is not cashless but it is consistent with using nominal variables without having to mention money balances. However, we also pointed out that additive separability is a very special, and arguably unrealistic, case.

Woodford considers the limiting *cashless* case for the more general model without additive separability. Thus money matters for real allocations and for price-level determination, but it matters less and less. For this approach we assume that the "relevance" of real money balances in the utility function (in comparison to consumption) can be represented by a parameter, ω, that can be taken to zero. Is it possible that for this limit in equilibrium (i) real money balances go to zero, while (ii) the price level remains finite and determinate and the remaining real variables can be pinned down as well? For this case consumers would demand less and less money from the central bank (in nominal terms), since they care less and less about money as ω approaches zero. Let us look into this possibility here.

Thus, consider the Euler equation for nominal bonds,

$$1 + i_t = \beta^{-1} \frac{u_{1,t}(c_t, M_{t+1}/P_t)}{u_{1,t+1}(c_{t+1}, M_{t+2}/P_{t+1})} \frac{P_{t+1}}{P_t},$$

which combines the Euler equation for real bonds, equations (17.31) and (17.33), with the Fisher equation, together with the money demand equation (17.35),

$$\frac{u_{2,t}(c_t, M_{t+1}/P_t)}{u_{1,t}(c_t, M_{t+1}/P_t)} = \frac{i_t}{1 + i_t}.$$

These two equations are interdependent but let us consider the first equation only in a limiting case. Let us again focus on an interest rate rule with a constant inflation target, $1 + \widetilde{\pi} = \widetilde{P}_{t+1}/\widetilde{P}_t$, and implicit interest rate target, $1 + \widetilde{i} = 1 + \phi(1) = (1 + r)(1 + \widetilde{\pi})$. With constant consumption in steady state, we have $1 + r = 1/\beta$. The idea is now to examine local determinacy around steady states, with successively smaller ω's, i.e., with lower and

lower values for real money balances in steady state, with a particular focus on the limiting case of zero. We will define the local dynamics in logarithms; the interpretation is thus one of how percentage changes in variables relate to each other, which is a robust notion also when some variables are close to zero. We thus define

$$\widehat{i}_t \equiv \log \frac{1+i_t}{1+\widetilde{i}}, \quad \widehat{c}_t \equiv \log \frac{c_t}{\bar{c}}, \quad \widehat{m}_{1,t} \equiv \log \frac{M_{t+1}/P_t}{\bar{m}_1}, \quad \text{and} \quad \widehat{\pi}_{t+1} \equiv \log \frac{P_{t+1}/P_t}{1+\bar{\pi}}$$

where bars denote the steady states of variables associated with the policy implied interest rate $\widetilde{i}$. To make clear that our definition of real balances purchased at time t, $m_{1,t} = M_{t+1}/P_t$, differs from our previous definition of beginning of period real balances, $m_t = M_t/P_t$, we have added the subscript one.

We obtain the first-order Taylor expansion of the Euler equation,

$$\widehat{i}_t = \eta_{u_1,c}\left(\widehat{c}_{t+1} - \widehat{c}_t\right) + \eta_{u_1,m_1}\left(\widehat{m}_{1,t+1} - \widehat{m}_{1,t}\right) + \widehat{\pi}_{t+1},$$

where

$$\eta_{u_1,c} \equiv -\frac{\partial u_1}{\partial c}\frac{c}{u_1}$$

and

$$\eta_{u_1,m_1} \equiv -\frac{\partial u_1}{\partial m_1}\frac{m_1}{u_1}.$$

The idea is now to parameterize the utility function with an ω such that, when ω approaches zero, the associated steady-state $\bar{m}_1$ value goes to zero, η_{u_1,m_1} goes to zero, and $\eta_{u_1,c}$ goes to a finite value. If so, the dynamics implied by the linearized Euler equation do not involve money at all, so together with the interest rate rule, it defines a complete system.

A complete characterization of the class of utility functions satisfying these requirements is beyond the scope of the treatment here. An example, however, can be provided. So suppose that u is an increasing, strictly concave function, f, of a constant returns-to-scale CES index, h, in the two arguments c and m_1

$$h(c, m_1) = \left[(1-\omega)c^\rho + \omega m_1^\rho\right]^{1/\rho} \text{ with } \rho < 1.$$

Here our key parameter appears: ω is the weight on real balances. With this formulation, the money demand equation, $u_2(c, m_1)/u_1(c, m_1) = h_{m_1}/h_c = i/(1+i)$, becomes

$$\frac{M_{t+1}}{P_t} = c_t \left(\frac{1-\omega}{\omega} \cdot \frac{i_t}{1+i_t}\right)^{\frac{1}{\rho-1}}. \tag{17.41}$$

This implies a money demand which has unitary elasticity with respect to consumption and a constant (negative) elasticity with respect to the cost of holding money, $i/(1+i)$. Since $\rho < 1$ the demand for real balances is a decreasing function of the opportunity cost of holding money. These features hold regardless of the value of ω: if $\omega \to 0$, then $M_{t+1}/P_t \to 0$ but the elasticities remain unchanged. Thus, if P_t is targeted to equal $\widetilde{P}_t$, with the aid of the interest rate rule, it must be that $M_{t+1} \to 0$.[17] It remains to be shown that when ω approaches

[17] For a more general utility function, we can rewrite the implicit money demand equation as

$$\frac{i}{1+i}\frac{m_1}{c} = \frac{u_2(c, m_1)}{u_1(c, m_1)}\frac{m_1}{c} = \frac{u_{1,2}m_1/u_1}{u_{2,1}c/u_2} = \frac{\eta_{u_1,m_1}}{\eta_{u_2,c}}$$

zero, η_{u_1,m_1} goes to zero while $\eta_{u_1,c}$ goes to a positive constant. It is straightforward, but somewhat tedious, to show this to be true under the functional assumptions given.[18]

Notice that the interest-rate rule is critical for the logic of the cashless limit to go through: the interest rate is fixed in a relation to prices such that, as ω goes to zero, M/P goes to zero through M going to zero. If monetary policy was instead governed by a money supply rule—the simplest form of which is to have M constant over time—then P would go to infinity (and p_m to zero) as ω and M/P went to zero, and the New Keynesian model could not be built on a core where the price level is infinity.

Monetary-fiscal interactions

In this section we look more carefully at the government's budget constraint and how it interacts with monetary policy. This will allow us to touch on a number of conceptual issues. Throughout, we maintain a consolidated view of the government sector, i.e., the fiscal authority plus the central bank. We do not model their behavior but rather discuss the set of policies they could undertake. An alternative would be to formulate a game between two authorities with different objective functions, and a private sector responding to policy, but there is no established approach to that in the literature.

Financing a given path of primary deficits In real terms, the flow budget constraint of the government reads

$$g_t + (1 + i_{t-1})\frac{B_t}{P_t} = \tau_t + \frac{B_{t+1}}{P_t} + \frac{M_{t+1} - M_t}{P_t}. \tag{17.42}$$

Here, the left-hand side is government spending—real purchases of goods plus debt repayment, including interest—and the right-hand side describes how spending is financed: via taxes, new borrowing, or money printing, also known as seigniorage. This budget constraint consolidates the budgets of the fiscal and monetary authorities; in particular, B denotes debt to the public, and any debts between the fiscal and monetary authorities net to zero. In equilibrium, the present value of the budget constraint reads

$$\sum_{t=0}^{\infty} q_{0,t}\,(g_t - \tau_t) + (1 + i_{-1})\frac{B_0}{P_0} = \sum_{t=0}^{\infty} q_{0,t}\frac{M_{t+1} - M_t}{P_t}, \tag{17.43}$$

where discounting, $q_{0,t} = \prod_{s=0}^{t} q_s = 1/\prod_{s=0}^{t}(1 + r_s)$, uses the Fisher equation.[19] The left-hand side of this equation is the present value of the primary deficit, $g - \tau$, plus the initial

since second derivatives (represented by $u_{i,j}$) are symmetric, $u_{1,2} = u_{2,1}$. A more general condition that delivers $m_1/c \to 0$ under a constant interest rate is therefore that $\eta_{u_1,m_1}/\eta_{u_2,c} \to 0$ as $m_1/p \to 0$, i.e., that the ratio of cross-elasticities of marginal utility goes to zero.

[18]Note, that $u_{1,1}c/u_1 = (f''c/f')h_1 + h_{1,1}c/h_1$ since $u_1 = f'h_1$. It is easy to see, by taking successive derivatives of the CES function h, that in the limit as $\omega \to 0$, the second term $h_{1,1}c/h_1 \to 0$. However, the first term becomes $f''(c)c/f'(c)$, which is strictly negative; hence, $\eta_{u_1,c}$ is strictly positive. Turning to $u_{1,2}m_1/u_1$, we obtain $(f''/f')h_{2,2} + h_{1,2}m_1/h_1$. As $\omega \to 0$, both these terms go to zero: the h function has the property that both $h_{m_1m_1}$ and $h_{1,2}m_1/h_1$ go to zero. Hence η_{u_1,m_1} goes to zero.

[19]One might think that the government's present value budget constraint represents the forward solution of the flow budget constraint (17.42), subject to a no-Ponzi game condition, analogous to the household's

debt (including interest), and the right-hand side indicates the total remaining financing needs that will have to be covered by money printing. In this obvious sense, the central bank's operations play a role in the consolidated budget.

There is also an alternative description of the consolidated constraint, which we can obtain by defining $D_t \equiv (1+i_{t-1})B_t+M_t$ as the total liabilities of the consolidated government at time t. Here, M is not a liability in the usual sense of something having to be paid back, but rather it is an item on the central bank's balance sheet.[20] Using this definition, the flow budget constraint can be written (along the lines of the description of the household's constraint, equation (17.11)) as

$$d_t = \tau_t - g_t + \frac{M_{t+1}}{P_t}\frac{i_t}{1+i_t} + \frac{1}{1+r_t}\cdot d_{t+1}. \tag{17.44}$$

Thus, the given initial liabilities $d = D/P$, which can be seen as a present value, equal the current primary surplus, $\tau - g$, plus the seigniorage revenue from the current presence of money in D, plus the present value of the liabilities left for next period, d'. This notion of seigniorage is more narrow and it represents a form of arbitrage that the government carries out: to the extent the private sector values money for non-financial reasons, the government saves on its financing costs by substituting non-interest-bearing money, M, for interest-bearing debt, B, while maintaining the same value for total liabilities, D. The interest savings, per unit of money, is $i_t/(1+i_t)$, and the total savings are larger, the larger the real balances share of the liabilities. We can also consider this form of seigniorage as that accruing for a fixed path of D/P.

Equilibrium seigniorage in the long run and fiscal dominance The previous discussion merely describes government variables, assuming that in equilibrium money and bonds are valued by private agents in the amounts specified. The value of money to the private sector has been the subject of previous sections. In the overlapping-generations setting, money serves as a mere storage instrument and only has value ($P < \infty$) if nominal interest rates are zero. Let us instead adopt the reduced-form liquidity perspective where we can derive a static money demand equation as $M_{t+1}/P_t = c_t f(i_t)$, with c being private-sector consumption and f a strictly decreasing function.[21] Thus, taking c_t as given, the arbitrage-based seigniorage expressed in (17.44) is proportional to $f(i)i/(1+i)$. This function has a decreasing part, $f(i)$, multiplying an increasing part, $i/(1+i)$, that starts at 0 and is bounded above by 1. So, under rather weak assumptions on f, the product expression will have a maximum and describe a "Laffer curve." As the nominal interest rate i increases above zero, seigniorage revenues are obtained, but at some point, further increases in i will lead to declining revenues.

budget constraint in Chapter 4, Section 4.3.1. There is, however, no optimization problem associated with the government, and thus there is no reason to impose an nPg condition on the government. Rather, we derive the government's present value budget constraint from the household's present value budget constraint by imposing market clearing. In this sense, the present value budget constraint (17.43) represents both a constraint and an equilibrium condition.

[20] In the distant past, when some central banks promised that each unit of money could be exchanged for a fixed number of units of gold, the liability was more apparent.

[21] The unitary elasticity in c obtains under special functional-form assumptions only but is not key to the argument here.

Let us now consider a stationary, long-run equilibrium: taxes and spending are constant, $\bar{\tau}$ and $\bar{g}$, as is consumption, $\bar{c}$, inflation, $\bar{\pi} = P'/P$, the real and nominal interest rate, $\bar{q} = 1/(1+\bar{r})$ and $\bar{i}$, and the real liability and money demand levels, $\bar{d}$ and $\bar{m}$. Then, the government budget constraint (17.44) will read

$$\bar{g} + \frac{\bar{r}}{1+\bar{r}}\bar{d} = \bar{\tau} + \bar{c}f(\bar{i})\frac{\bar{i}}{1+\bar{i}}.$$

Many of the models in the literature have exact, or approximate, separation between purely monetary and real phenomena in the long run. With that motivation, let us consider $\bar{g}$, $\bar{r}$, and $\bar{c}$ to be given here. Then an increase in government liabilities (be it through an increase in money or bonds) must cause an adjustment in the nominal interest rate. If, in addition, the economy is on the "benign" side of the Laffer curve, then an increased $\bar{d}$ will raise $\bar{i}$. Because $\bar{r}$ is given, the Fisher equation implies that inflation will need to rise.

The long-run version of the government's budget constraint allows us to touch on the well-known Sargent and Wallace (1985) piece "Some Unpleasant Monetarist Arithmetic." In that paper, the authors argue that if the central bank at any point wishes to fight inflation through an open-market operation where the money stock is decreased and the stock of debt is increased (borrowing to buy and withdraw money from the economy), then because debt pays interest, $\bar{d}$ will rise. So in the long run, $\bar{d}$ will be higher (unless the short-run action is reversed), which will increase $\bar{i}$ and, therefore, inflation. Thus, the arithmetic of the consolidated government's budget makes fighting inflation difficult if the fiscal authority does not collaborate by increasing its primary surplus. Sargent and Wallace framed their argument as one with a "fiscal dominance regime." That is, the fiscal authority independently commits to primary surpluses and leaves it to the monetary authority to adjust as necessary to satisfy the government's intertemporal budget constraint. In such a situation, even a monetarist central bank may not be able to fight inflation. Their argument can be made without restricting the analysis to a steady state, but we omit the details here.

The irrelevance of open-market operations Wallace (1981b) considers a situation where open-market operations, that is, changes in the central bank's balance sheet involving exchanges of money for outstanding government bonds, do not affect the economy at all. This occurs in a special situation where money functions as a store of value and is not dominated in return by government bonds. The consolidated budget setting here allows us to explain the irrelevance; the original argument in Wallace (1981) instead used the overlapping-generations model of money to derive the result.

Recall again the budget constraint (17.44) and suppose now that $i_t = 0$ for all t. The budget then simply reads

$$\frac{M_t + B_t}{P_t} = \tau_t - g_t + \frac{1}{r_t}\frac{M_{t+1} + B_{t+1}}{P_{t+1}}.$$

It is not clear from this budget constraint whether (and why) money has value, but the maintained assumption is that the private sector regards M and B as perfect substitutes. In addition, it is not clear how many equilibria may exist. However, the following can be said: if an equilibrium exists, in the form of a set of sequences $\{M_t, B_t, \tau_t, g_t, r_t\}_{t=0}^{\infty}$ along with other

variables (consumption, output, relative prices, etc.), then any $\left\{\widehat{M}_t, \widehat{B}_t, \tau_t, g_t, r_t\right\}_{t=0}^{\infty}$ with other variables maintaining their values, is also an equilibrium so long as $\widehat{M}_t + \widehat{B}_t = M_t + B_t$ for all t. The proof is trivial: the government's budget constraint is still met and the private sector's situation is entirely unchanged, so if the original sequences constitute an equilibrium, so does the proposed alternative.

All in all, the only truly central assumption behind the irrelevance proposition is that money and bonds be perfect substitutes. This assumption might make you think that the proposition itself is irrelevant: in reality, bonds pay interest and money does not! However, the recent long period of zero interest rates again made the proposition relevant: it suggests that *quantitative easing*, QE, is irrelevant at the zero lower bound. For QE to have real impact, bonds and money have to differ for some reason. Most discussions of QE focus on differences in duration—central banks exchange money for long-term bonds. The irrelevance proposition thus forces us to think more about how money and bonds differ.

Indeterminacy issues: fiscal back stops and the fiscal theory of the price level We have seen, in the various discussions of different models of money, that monetary equilibria may not be unique. Moreover, the number of equilibria may depend on the monetary policy assumed. We now consider two examples. In the first example, a "fiscal backstop" eliminates all equilibria for which money loses value in the limit. In the second example, illustrating the *fiscal theory of the price level*, a policy that fixes the present value of government net-revenues eliminates a continuum of equilibria that start away from the stationary equilibrium but eventually converge to it. In either case, we are left with the unique stationary equilibrium.

So, first, recall the hyperinflation example above with a continuum of equilibria indexed by the initial price level P_0, where the equilibrium price path increases without bounds for any initial $P_0 > \bar{P}_0$ in such a way that money loses value entirely in the limit. The fiscal-monetary policy specification for this example was constant money growth $M_t = \gamma M_{t-1}$ implemented through lump-sum transfers, that is, negative taxes. A simple fiscal backstop that eliminates the equilibria with price level growth exceeding money growth is one where the government promises to buy an unlimited amount of money if, at any time, the price level exceeds a critical value, $\widetilde{P}_t = \widetilde{P}_0 \gamma^t$, with the purchases financed by lump-sum taxes; see Wallace (1981a) and Obstfeld and Rogoff (1983). The critical value grows at the rate γ, since in the stationary equilibrium with constant real balances the price level grows at the money growth rate. If the market considers this policy credible, the hyperinflationary equilibria cease to exist. Thus, fiscal policy supports a monetary equilibrium through an off-equilibrium action.[22]

Now consider another example where fiscal policy is used to rule out indeterminacy of the equilibrium path. In particular, consider the pure interest-rate peg in the context of the general-equilibrium cash-in-advance model. For this policy the initial price level, P_0, is indeterminate and indexes all perfect foresight equilibria. Furthermore, the interest rate peg is supported by a path for the nominal money stock, M_{t+1}. We now describe a policy that eliminates equilibrium indeterminacy. We will consider a specific policy for simplicity, but the argument is general in nature: it only relies on the government's budget constraint and

[22]Notice that this fiscal-monetary policy effectively amounts to a promise to give money value in the future, that is, money is "backed" (by goods, not gold).

an assumption about what variables adjust to make sure it is satisfied. Hence, the argument applies for a general class of models displaying indeterminacy. We will consider a simple rule for lump-sum taxes

$$\tau_t - \tilde{\tau} = -\frac{M_{t+1} - M_t}{P_t} \tag{17.45}$$

for a given $\tilde{\tau}$ and fixed spending $g_t = \tilde{g}$. Here, lump-sum taxes decline one-for-one with the magnitude of the real money stock change necessary to support the interest rate peg on the equilibrium path. Thus, if seigniorage were to decline, so that the right-hand side would rise, the lump-sum tax would go up to compensate for the seigniorage shortfall. Substituting the tax rule into the government's budget constraint (17.43) yields

$$\frac{(1+i_{-1})B_0}{P_0} = \sum_{t=0}^{\infty} \beta^t \left(\tau_t - g_t + \frac{M_{t+1} - M_t}{P_t}\right) = \sum_{t=0}^{\infty} \beta^t \left(\tilde{\tau} - \tilde{g}\right) = \frac{\tilde{\tau} - \tilde{g}}{1-\beta} \tag{17.46}$$

Here, the only endogenous variable is the price level P_0, and the promise of future fiscal adjustments thus implies a unique solution for the price level. The *fiscal theory of the price level* refers to the notion that commitment to a fiscal rule for the indefinite future determines the present value of primary surpluses, which forces the price level to adjust so that the initial real debt is covered by the present value of future surpluses. The argument of course requires that $B_0 > 0$, but the other details of our example are not important. One can view government debt here as an asset: a (forward-looking) claim to future payments in the form of primary surpluses.

Compare the outcome for this fiscal rule with the outcome for an interest rate peg we studied above when the interest rate is determined by a state contingent rule, equation (17.39). There we have shown that if the interest rate rule is sufficiently responsive to deviations of the price level from a target the equilibrium is also unique. But if the price level is unique, how can we the sure that it is consistent with the price level as determined by the government's present value budget constraint? Here it becomes important that the present value government budget constraint represents an equilibrium outcome. Therefore if the nominal interest rate rule (17.39) results in a unique equilibrium with associated paths for the price level and money stock then fiscal policy, lump sum taxes and spending, has to adjust such that the present value budget constraint holds. Again, the problem is related to the Sargent and Wallace (1985) piece "Some Unpleasant Monetarist Arithmetic": the price level being uniquely determined by the interest rate rule implicitly assumes a "monetary dominance" regime.

There is debate among macroeconomists about the relevance of the fiscal theory of the price level. To some, it seems counterintuitive that price setters today—imagine a fruit seller—are modeled as needing to set the price so as to make sure that the value of the government's debt becomes equal to a certain value. But note that we are talking about equilibrium prices in a dynamic competitive equilibrium, and in general, we study the properties of an equilibrium and do not ask how an equilibrium comes about.

17.4 Multiple currencies

What determines the exchange rate between two currencies, such as the dollar and the euro? This is a major question in international macroeconomics. Undergraduate textbook treatments refer to certain "parity" conditions—purchasing power parity or (covered or uncovered) interest rate parity—as guidance. Here, the purpose is to discuss exchange rate determination from the perspective of the models considered so far. In so doing, we will also more generally touch on implications for determining the value of other money-like assets, such as crypto-currency.

17.4.1 Money as a store of value: Kareken-Wallace exchange rate indeterminacy

What if we introduced more than one type of money (or currency) into models where money's only role is that of a store of value, such as the overlapping-generations model? The main discussion here will be in the context of one country only. One reason for this is that it simplifies the exposition. Another is that consumer heterogeneity or multiple types of goods is not central to the argument. We will also consider the possibility that currency stocks grow at different rates, mimicking differences in monetary policies across countries. The discussion follows Kareken and Wallace (1981).

So, consider an overlapping-generations model with one good, one representative consumer per cohort, and two currencies, a and b, for the consumer to invest in. The prices of the two currencies in terms of the consumption good are p_a and p_b. The consumer born at t thus faces budget constraints

$$c_y + p_{a,t}M'_a + p_{b,t}M'_b = \omega_y \quad \text{and} \quad c_o = \omega_o + p_{a,t+1}M'_a + p_{b,t+1}M'_b,$$

and the non-negativity constraints, $M'_a \geq 0$ and $M'_b \geq 0$. The choice is over (c_y, M'_a, M'_b, c_o) and let us for simplicity use $u(c_y, c_o) = \log c_y + \log c_o$ as the objective function to maximize. Here, the consumer faces four prices but what matters for decision-making are the real returns $p_{a,t+1}/p_{a,t}$ and $p_{b,t+1}/p_{b,t}$. Because short-selling is not possible, the consumer cannot conduct arbitrage based on rate of return differences between the currencies. Thus, if at any point in time t one currency has a lower return than the other currency, no consumer will hold the low-return currency: its price will be zero. It then follows that the price would be zero also before that date since no consumer would buy the currency for a positive price and sell it later for a zero price. Similarly, the price of this currency after the date t would also have to be zero since otherwise, markets would not clear in the period before the price becomes positive: consumers would demand an infinite amount of currency at that point. Hence possible equilibria have the structure that at all times, either (i) both currencies are valued; (ii) only one of the currencies is valued; or (iii) no currency is valued. As the discussion in Section 17.2 should make clear, case (iii) will apply if $\omega_y \leq \omega_o$. If $\omega_y > \omega_o$, however, (i) and (ii) are possible. But case (ii) is subsumed in our previous analysis. If, for example, future agents do not value currency a, present agents will not either, and hence $p_{a,t} = 0$ at all times and we are back to the overlapping-generations economy with one currency. So instead consider case (i) and let e_t be the *nominal exchange rate* between the currencies,

$e_t \equiv P_{b,t}/P_{a,t} = p_{a,t}/p_{b,t}$. The exchange rate measures how many units of currency b need to be given up to obtain one unit of currency a: if e is above 1, one unit of currency a is more valuable than one unit of currency b. Clearly, from $p_{a,t+1}/p_{a,t} = p_{b,t+1}/p_{b,t}$ at all times, it follows that e_t must be constant over time: $e_t = e$. Moreover, we can define the total money holdings, in currency b units, as $M' \equiv eM'_a + M'_b$ and $p_b M'$ as its real value. Thus, the consumer's budget constraints become

$$c_y + p_{b,t}M' = \omega_y \quad \text{and} \quad c_o = \omega_o + p_{b,t+1}M',$$

and, just as in Section 17.2, demand for total money by the consumer born at t satisfies the equation $p_{b,t}M_{t+1} = \max\{(\omega_y - \omega_o p_{b,t}/p_{b,t+1})/2, 0\}$. To close the model, we need to specify money supplies. Suppose they are constant over time: for all t, $M_{a,t} = M_a$ and $M_{b,t} = M_b$. Thus, focusing on equilibria where money has value, $M = eM_a + M_b$ and the equilibrium is determined by

$$p_{b,t}M = \frac{1}{2}\left(\omega_y - \frac{\omega_o}{p_{b,t+1}/p_{b,t}}\right)$$

holding at all times. As in the one-currency model, this defines a set of equilibrium sequences for $p_{b,t}M$: one is a steady state, where $p_{b,t}M$ is constant and equal to $(\omega_y - \omega_o)/2$; but there is also a continuum of other paths with $p_{b,t}M$ converging to zero over time. The key observation here, however, is that there is no other equilibrium condition, and hence e is not determined. To be more concrete: select an arbitrary $e \in (0, \infty)$. This will define $M = eM_a + M_b$, since (M_a, M_b) are given. Then $p_{b,t}$ follows from knowing $p_{b,t}M$ in the given equilibrium. This allows us to find $p_{a,t}$: it equals $ep_{b,t}$ at all times. Since the consumer's demands for individual currencies are not pinned down—there is complete indifference, since the currencies give identical returns—we can then set $M_{a,t+1} = M_a$ and $M_{bt+1} = M_b$ at all times, since their value sum is then equal to their chosen total amount of saving $(eM_a + M_b)p_{b,t}$.

Suppose now that the currency stocks change over time because the government provides lump-sum transfers of the two currencies to the agents when old. We can still, for an arbitrary e, define a total money supply as $M_t = eM_{a,t} + M_{b,t}$, which will now in general depend on time. Incorporating the lump-sum transfers and after some algebra we derive a modified demand for money that again defines a difference equation for real balances

$$p_{b,t}M_t = \frac{1}{2}\left\{\omega_y - \frac{\omega_o + p_{b,t+1}(\gamma_t - 1)M_t}{p_{b,t+1}/p_{b,t}}\right\},$$

where $\gamma_t = M_{t+1}/M_t$. A solution will be nonstationary because γ_t appears in the equation. Consider the simple case where each money stock grows at a constant gross rate, γ_a and γ_b, and without loss of generality, suppose $\gamma_a \geq \gamma_b$. Then γ_t will converge to γ_a. One can show that an equilibrium exists where total real balances converge to a positive constant with the limit being $p_b M = (\omega_y - \gamma_a \omega_o)/(1 + \gamma_a)$.[23] Thus faster-growing money will dictate the long-run real value of the total money stock, and it will comprise all of the money stock, $eM_a/M = 1$. This can be interpreted as a version of Gresham's law: "bad money drives out good money," where "bad" refers to a faster-growing stock. This statement, of course, is

[23] Since the proof is a bit complicated we leave it to the reader to consult the Appendix of Kareken and Wallace (1981).

conditioned on an equilibrium of type (i); another equilibrium is always that where $e = 0$, in which the value of the total money stock will converge to $p_b M = (\omega_y - \gamma_b \omega_o) / (1 + \gamma_b)$, which is higher.

17.4.2 Dynastic models with a reduced-form liquidity demand

With a reduced-form liquidity demand, a key question immediately arises: how should the two currencies appear (in cash-in-advance constraints, in a transactions-cost technology, or in utility)? In an early paper, Lucas (1982) considers a two-country model where there are two traded goods and consumers in both countries value both goods. Country 1 consumers, however, are only endowed with goods of type 1 and country 2 consumers are only endowed with goods of type 2. He, then, assumes that good 1 is subject to a cash-in-advance constraint involving only country 1 money, and similarly for good 2: you need country 2 money to buy it. Therefore, all consumers need both types of currency. The model gives a uniquely determined exchange rate: the value of country 1 money is tied, via the cash-in-advance constraint, to the total demand for good 1, which is real. Thus, the exchange rate is directly tied to the relative demands for the two consumption goods (and to the relative money stocks).

A similar result to Lucas's can be obtained with any of the other models where money has reduced-form liquidity demand. One can, for example, assume that foreign money has some (quantitatively limited) value to one's utility. Monetary policies will matter for exchange rates since the real value of money is pinned down by its reduced-form real role, as well as its financial return. If a country increases its money stock at a slower rate than other countries, then its exchange rate will appreciate over time, everything else equal.

Discussion: theory and data

Exchange rates fluctuate significantly over time and are typically described as random walks, that is, their movements are unpredictable. To what extent can the theories used here be used to understand these facts? The overlapping-generations model we studied above does not generate fluctuations: it predicts an indeterminate but constant exchange rate. However, for extensions of the model that introduce extrinsic uncertainty (i.e., non-fundamental random fluctuations, such as "sunspots") the indeterminacy then allows for unpredictable movements in exchange rates (martingales). Reduced-form liquidity models admit randomness in exchange rates to the extent that there is randomness in fundamentals, such as in money supplies or output.

Relatedly, practitioners (and basic undergraduate textbooks) often refer to *Purchasing Power Parity* (PPP) as a guide for understanding what the value of an exchange rate should be: it should cost the same to buy a given set of (tradable) goods in one currency directly as it would if one swapped into another currency and bought the goods using that currency.[a] Domestic prices do not fluctuate nearly as much as do exchange rates, implying that PPP cannot hold at all points in time, even though it might hold on average over time. So to the extent one could identify a basket of goods that is available in two countries, couldn't this be a way to think about exchange rate deter-

mination? Clearly, in the models described above—the overlapping-generations model and the reduced-form liquidity models—PPP holds: there is free trade in goods and currencies. If purchasing power parity appears to be violated in the data, then in a strict sense it contradicts these theories. But the idea here would be to argue that the theories still hold on average over time, and hence they can be used to predict an upcoming adjustment in exchange rates in the direction of making PPP hold. However, PPP can be restored also by adjustments in the price levels: these are fundamentally endogenous. Prices may move slowly over time, but gradual movements would also allow us to move toward restoring PPP, thus not necessarily involving exchange rate adjustments. A similar condition is *interest rate parity*: the notion that investing in bonds in one country should give the same return as investing in bonds in another country. In particular, a dollar invested in U.S. bonds gives an interest of i_t between t and $t+1$; alternatively, the investor could buy euro at the time t exchange rate, invest in euro bonds paying $\tilde{i}_t$, and convert the proceeds back to dollars at the $t+1$ exchange rate. Again, in all the theories considered above, these two transactions would give the same return. Can apparent departures from this kind of parity be used to argue that the current exchange rate is too high or too low? No. Parity does not predict the exchange rate at t (conditional on knowing i_t and $\tilde{i}_t$), rather it predicts the exchange rate at $t+1$ relative to that at t. Hence, it cannot help us understand the level of an exchange rate.

Because of their large fluctuations and violation of parity conditions, exchange rate movements are challenging to understand with our basic theories. Exchange rate indeterminacy in the overlapping-generations model suggests a way of thinking about fluctuations. However, it does not account for why dollars are used in most transactions in the United States and euro in euro countries. Reduced-form models, with the stroke of a pen, allow for this feature. They also predict that if a country prints money at a faster rate than another country, and it experiences higher inflation as a result, then its exchange rate will depreciate. This feature is broadly in line with data. One interesting example is the Swiss franc, a currency that has experienced decades of consistent appreciation. Is this appreciation because Swiss inflation has been low in an international comparison? Maybe yes, but the appreciation is much stronger than what can be accounted for (using a PPP relationship) by its low inflation. Thus, the obvious fundamentals go some, but far from all of, the way toward understanding the facts.

[a]The so-called Big Mac index is another example of this idea, though a Big Mac can hardly be regarded as tradable.

17.4.3 Crypto-currency

Crypto-currency is a form of digital currency that can serve as a medium of exchange and a store of value. The circulation of alternative currencies has a long history, for example, paper money issued by commercial banks. Whether the digital nature of crypto-currency makes it special is not clear but it does have certain distinguishable properties: it is costless to "carry" and, typically, has a way of securing privacy in that your holdings are, for example, not directly observable to the government. Thus, crypto-currency may have a comparative advantage for criminal activities, especially since regulations limiting money laundering have

become more and more potent in many countries. But crypto-currency can be given different formats too; in some versions, it is marketed as an asset simply with above-market return (for some time), and in others as a safe store of value (for example, the so-called stable-coins, whose value is tied to the dollar).

How can crypto-currency be understood given the above theories? The overlapping-generations model would say that it is just another currency that can, based on the expectations of the behavior of future consumers, potentially be used as a store of value. Thus, its value while indeterminate could be positive when introduced. And its exchange rate with standard currencies could move randomly. This theory does not rely on any of the specific properties of crypto-currency. The reduced-form models of liquidity could also accommodate more currencies. A cash-in-advance theory could allow other currencies as allowable means of payment for some or all of the goods. A money-in-the-utility function model could introduce another variable in utility, perhaps nested with standard money in a CES form. Of course, none of these approaches seem satisfactory since the results (in the form of the equilibrium value of crypto-currency) appear to follow very directly from the assumptions.

17.5 Missing assets

In the remaining sections of this chapter, we look at two more fundamental models of money in which valued money is not an assumption but is an outcome of the environment. As we have seen, in the overlapping-generations model, money can serve as a store of value under some restrictions on primitives, primarily involving utility functions and the time profile of endowments. In the present section, we look at an environment in which money as a store of value can (at least partially) replace some missing assets. We introduce this feature into a dynastic environment closely related to material covered earlier in the book: consumer heterogeneity and incomplete insurance against idiosyncratic shocks.

Consider the Huggett model studied in Chapter 11, with the tightest possible borrowing constraint: agents cannot borrow, $a' \geq 0$. In a stationary equilibrium agents solve

$$v_\omega(a) = \max_{a' \geq 0} u(a + \omega - qa') + \beta \int_{\omega'} v_{\omega'}(a') F(d\omega'|\omega),$$

where income ω follows a first-order Markov process. We obtain an autarky equilibrium where nobody borrows or lends. The agent valuing the bond the most is indifferent between borrowing and lending, and this agent's Euler-equation determines the interest rate; everybody else is borrowing constrained. In the simplest, two-state case with nontrivial probabilities, we obtain

$$qu'(\omega_{hi}) = \beta \left[\pi_{hi|hi} u'(\omega_{hi}) + \pi_{lo|hi} u'(\omega_{lo}) \right]$$

and hence the gross real interest rate is

$$\frac{1}{q} = \frac{1}{\beta} \cdot \frac{1}{\pi_{hi|hi} + \pi_{lo|hi} \frac{u'(\omega_{lo})}{u'(\omega_{hi})}} < \frac{1}{\beta}.$$

Now suppose $1/q < 1$, that is, we have a negative (net) real interest rate, like in the overlapping-generations case. Here as well, people would like to save (rather badly if

$u'(\omega_{lo})/u'(\omega_{hi})$ is very high) but are lacking assets for it: intra-personal loans are not allowed because the asset cannot be held in negative amounts. That is, an asset is prevented from existing. Other assets are missing too—insurance claims written contingent on idiosyncratic endowment outcomes—but the key here is that a riskless asset is missing.

Now consider introducing fiat money in the economy: suppose there is a fixed stock of M nominal units. If the price level is constant, $P_t = P$, then money can play the role of a riskless asset, $m = a$, with gross real return one, $q = 1$, that helps consumers with high-endowment realizations save and smooth consumption.[24] What would the steady-state real value of this money be? The answer can be obtained by solving

$$v_{\omega}(a) = \max_{a' \geq 0} u(a + \omega - a') + \beta \int_{\omega'} v_{\omega'}(a') F(d\omega'|\omega);$$

with implied decision rule: $a' = g_{\omega}(a)$. Thus the total value of the money stock is simply the total savings in this economy

$$m = \int_{a \geq 0, \omega} g_{\omega}(a) \Gamma(d\omega, da),$$

where Γ is the stationary distribution implied by the decision rule and the distribution of endowment shocks. Given the fixed stock of nominal money, this expression then determines the price level. Of course, the role played by money could also be played by (real) government bonds.

An even simpler missing-asset model is a deterministic version of the model above where endowments alternate between high and low: for half of the population, endowments are high (low) in even (odd) periods, and for the other half of the population they are high (low) in odd (even) periods. If borrowing is not allowed (individual debt assets are ruled out), then the equilibrium would be autarky. But money could be introduced and would have value if the autarky interest rate in the original economy is below zero. This is essentially the environment of Townsend (1980)'s turnpike model of money. Like the overlapping-generations model, the missing-asset model has the attractive feature that it can account for valued money without simply assuming it, but it cannot explain why money continues to have value when being dominated in return. One might argue that even though money is not valued by assumption in these environments, money is valued because some other assets are excluded by assumption.

17.6 Models of money as a medium of exchange

The cash-in-advance and transaction cost models described above share the feature that money is used to facilitate consumption and, in that sense, functions as a medium of exchange. However, these are not models that explain why money may be important in exchange. In undergraduate texts, the informal motivation for money as a medium of exchange is the *absence of double coincidence of wants*. That is, when buyers meet sellers they are

[24] Similar to the overlapping-generations model, one can also imagine non-stationary equilibria here, where money would lose value over time and be worthless in the limit.

rarely in a situation where a direct exchange of goods or services is mutually beneficial (without access to some form of public record-keeping technology). The model in Kiyotaki and Wright (1989) is the first that carefully spells out the friction that motivates this idea. We briefly describe a setting like theirs here (without equations). A modern version of their model is contained in Lagos and Wright (2005). In all these settings, trade is fundamentally *decentralized*: people don't (always, at least) trade in centralized markets. Also, markets are *anonymous*: there is no record-keeping. In this sense, they are similar to the search/matching model in Diamond (1982).

Kiyotaki and Wright's (1989) model We will describe the core setting and attempt to explain verbally how it works. There are three kinds of people, all deriving utility from the consumption of a good, but a consumer of type $i \in \{1, 2, 3\}$ only consumes good $i+1$, modulo 3 (i.e., $i + 1$ is interpreted as 1 when $i = 3$). We denote the utility benefit of consumption by u_i for consumer i. Moreover, there is production: upon consuming, a person of type i produces a good of type i and, hence, is also a producer. Hence, a person's identity i is the good she produces. All goods must be indivisible: production, when it occurs, always delivers one unit of a good. Each period, people meet pairwise. Thus, by construction, when people meet, there is never a double coincidence of consumption wants. For example, a person of type 1 can offer her produced good 1, but only a person of type 3 likes that good, and unfortunately that person produces good 3, which person 1 does not consume.

In this economy, goods can be stored. In particular, good i can be stored from t to $t + 1$ at a utility cost c_i that is additive (does not interact with consumption utility). Thus, it is conceivable that a producer of type i, when meeting someone, swaps the good just produced for another good with the idea of storing it and possibly swapping this good for good $i+1$ and consuming it in the future. The stored good would then function as a *medium of exchange*. Meetings are assumed to be random and a person of type i is equally likely to meet a person of type $i + 1$ and $i + 2$, modulo 3.

In each period of the model, a person only has one choice: upon meeting another person, whether or not to swap goods, i.e., trade, as a function of what good the other person is carrying. A Nash equilibrium concept is used whereby a swap will only occur voluntarily, i.e., if both people's choices are to trade.

It is relatively straightforward to characterize the set of steady-state equilibria for this model. People's decisions will depend on an expectation of what other people will do. Hence, deciding to accept a good and store it only makes sense if one's expectation is that this good will be accepted by others and that such an exchange can (eventually) lead one to be able to consume.[25] Thus, the distribution of goods holdings in the population and the associated exchange strategies of agents is key for decision making today. Existence of equilibria is made easier if randomization is allowed, but pure-strategy equilibria can occur as well. Which good(s) might appear as a medium of exchange is a function, naturally, of the relative storage costs, but also of the relative utility benefits and of consumers' rate of discount. Similarly, whether a given good is a *general* medium of exchange (and thus is accepted in trade against all other goods) depends non-trivially on the parameters of the

[25] Here, for example, one possibility is that person 1 trades the just-produced good 1 to obtain good 3, stores good 3 and then manages to trade good 3 for good 2.

model. Several steady-state equilibria may also exist, which is not surprising due to the central role played by expectations.

Now *fiat money* would enter this economy as another good that a subset of agents is endowed with at time zero. "Fiat" means that it is intrinsically useless: no consumer derives utility from its consumption, or has any production benefits from it. Like other goods, money is also indivisible and can be carried from period to period (in no greater quantity than one) at a utility cost $c_\$$. In addition, the storage properties of money are attractive: $c_\$$ is low. There is thus a fixed amount of money in the economy that could, potentially, function as a medium of exchange. Kiyotaki and Wright show that there are assumptions on the primitive parameters such that money, indeed, functions as a medium of exchange. Naturally, there is also another equilibrium where money has no value: it is never accepted in exchange, since it is not expected to be used in exchange in the future.

Discussion The Kiyotaki-Wright model of a medium of exchange satisfies Wallace (1998)'s dictum: the patterns of exchange emerge endogenously and non-trivially, including the role of fiat money in exchange. In the absence of public record-keeping, whereby people could get "credit" from giving up a good and hence obtain a good without exchange in the future, money at least partially plays this role: if a person comes into a meeting holding money, it must mean that that person gave up a good for money at some point in the past. The model has shortcomings, to be sure, the indivisibility of goods and money being one (relative "prices" in exchange are therefore either 1, 0, or infinity). The full absence of (centralized) markets also makes it difficult to imagine how monetary policy would be conducted. However, the original paper was followed up by many others, gradually relaxing these strong assumptions. In particular, Lagos and Wright (2005) relax all the assumptions just mentioned, yet their model remains tractable due to special assumptions made on utility functions whereby the distribution of money holdings collapses: all agents carry the same amount of money into each period.[26] To go through that framework in detail is second-year material, however. For a textbook treatment of the Lagos and Wright (2005) model and the literature that uses that framework see Rocheteau and Nosal (2017). It is also interesting to note that whereas the New Keynesian model, to be described below in Chapter 18, has been generalized in the direction of consumer and firm heterogeneity as well as a labor market with search frictions, it has not been merged with the medium-of-exchange settings. At least in part, this is because the New Keynesian model focuses on (i) cashlessness, with the argument that cash is more and more rarely used, and (ii) sticky prices, which is not the focus of the medium-of-exchange literature.

[26]It is, however, difficult, using their setting, to rule out the use of bonds as a means of payment in the decentralized market. Hence the setting becomes close to a cash-in-advance model where bonds are ruled out as a means of payment by assumption.

Chapter 18

Nominal frictions and business cycles

Alisdair McKay and Morten Ravn

18.1 Introduction

The Great Depression was a key event in the development of macroeconomics as a field. To many observers, this episode revealed that market economies can perform very poorly as the mass unemployment of the Depression was a clear sign that something was very wrong. It was then natural to ask what led to this failure and what could be done to improve the situation.

The Keynesian school of thought emerged as a result of the Depression and provides a perspective on business cycle fluctuations of all magnitudes not just crises like the Depression. One core idea of Keynesian economics is that the productive factors of the economy may not be used at the optimal level. A second core idea in Keynesian economics is that the level of nominal demand influences the level of production. A central piece of most economic theory is that, in the *long run*, nominal values are unimportant and simply a choice of units. However, going back at least to Hume's 1752 essay *Of Interest*, it has also been recognized that changes in nominal demand such as a change in the money supply may have real effects, at least temporarily, because the process of adjusting prices is neither perfect nor immediate.

This chapter serves several purposes. We will review the evidence that (a) prices adjust only infrequently and do not immediately react to changes in market conditions and (b) changes in nominal variables (here nominal interest rates) affect real outcomes such as real output. We also present the modern incarnation of the Keynesian tradition in the form of the workhorse New Keynesian model. Using this model, we will explain why imperfect price adjustment leads to a role for nominal demand in determining real outcomes, why the market equilibrium can differ from the efficient level of production, and what types of policies can lead to better outcomes.

18.2 Empirical evidence on price rigidity

Here we will review the empirical evidence on price rigidity. We will concentrate on nominal rigidities in prices because this is also the focus of most of our theoretical analysis, but many of the same issues that apply to the prices of goods and services also apply to wages.

The research we review here examines data on the prices of individual goods as opposed to economy-wide price indices. We discuss aggregate evidence later in the chapter.

The literature on the adjustment of individual prices developed significantly since the early 2000's. In a line of important papers focused on the U.S., Bils and Klenow (2004b), Klenow and Kryvtsov (2008) and Nakamura and Steinsson (2008) exploited access to the survey-based micro data underlying the construction of the Consumer Price Index (CPI) to document a series of facts about goods-level price adjustments. Much of this work has now been extended to other countries.[1]

The data underlying the U.S. CPI is collected by the Bureau of Labor Statistics (BLS). The BLS surveys are carried out monthly or bi-monthly in 75 urban areas collecting price quotes from about 23,000 retail and service establishments for about 85,000 individual items. These individual items are then classified into about 300 goods categories known as entry level items (ELIs).

One feature of these data that has received particular attention is the frequency of price changes. Suppose a price has a probability $1-\theta$ of changing each month. The expected time between price adjustments is then

$$\sum_{t=1}^{\infty}(1-\theta)\theta^t t = \frac{\theta}{1-\theta}.$$

Thus, observing the length of time a price of a good remains at a certain level is informative about the frequency of price changes for this good. In the data, such price change frequencies depend on the category of goods. For instance, magazine prices tend to change very infrequently while the prices of energy and food items tend to be adjusted much more frequently. The mean and median price duration of all items in the CPI are 6.2 months and 3.4 months, respectively (see Klenow and Malin, 2010).[2] The very considerable difference between the mean and the medians indicates significant heterogeneity across categories.

One of the features of these data is that prices have memory—sellers often return the price of an item to a level they have set in the past. One key source of this memory is sales, i.e. temporary discounts. The key problem presented by sales is whether such episodes should be excluded or not when estimating the frequency of price changes. If firms introduce sales as a direct response to a temporary change in market conditions, it would seem appropriate to include sales. In contrast, if the timing and size of the temporary price discount are not affected by economic conditions, price changes associated with sales may not reflect true flexibility in prices and should therefore be excluded. The literature therefore typically report estimates with or without controlling for sales.

Many individual products change over time and are replaced by new versions. Such product replacements imply that for a subset of goods, while their prices have been observed in the past, prices cannot be observed this month (and future months) because they are no longer for sale. In constructing the CPI, the BLS will substitute a new comparable product

[1] In this respect, an important effort has been made by the European Central Bank in its Inflation Persistence Network which has been summarized in Dhyne, Alvarez, Le Bihan, Veronese, Dias, Hoffmann, Jonker, Lunnemann, Rumler, and Vilmunen (2006).

[2] The statistics reported here are computed from the price change frequencies within ELIs weighted using CPI weights. These results relate to survey prices in the three largest cities in the U.S.

for the unavailable product. These product substitutions are often associated with price adjustments. It is an open question whether and how one should control for such product substitutions when computing measures of price rigidity. To the extent that product turnover presents an opportunity to adjust prices in the face of shocks, it would seem natural to include such price adjustments in measures meant to inform about the level of price rigidities. At the same time, by its very nature, when there is product substitution, the precise nature of the good changes and so the price change may be unrelated to market conditions.

When temporary sales prices are excluded, the estimates of the mean and median price durations rise to 8.0 months and 6.9 months, respectively (see Klenow and Malin, 2010). Hence, it is clear that whether sales are included or not makes a big difference especially if one concentrates on the median price duration estimates. These estimates treat price quotes as referring to the same item even in the face of product substitutions. Eliminating these as well, the mean and median price durations increase further to 10.1 and 8.3 months, respectively.

In summary, judging from the evidence across goods, there is considerable evidence of infrequent price changes with the range of empirically plausible estimates of price durations for the U.S. ranging from 6 months to 11 months depending on whether one focuses on the mean or the median and depending on the price measure. Estimates for European economies are typically in the upper range of this interval (see Dhyne et al., 2006).

What is not clear from this evidence is whether prices change infrequently because there are infrequent shocks to "market conditions" or because there are frictions that prevent prices from adjusting. Eichenbaum, Jaimovich, and Rebelo (2011) use data from a large retailer to estimate how the prices of products respond to the replacement cost of the goods and find that the price-cost margin varies within a fairly narrow range implying prices respond quickly to changes in market conditions.

Movements in individual prices are large relative the changes in the aggregate price level. For example, the median absolute price change is 11.5 percent (see Klenow and Kryvtsov, 2008). The obvious interpretation of this fact is that most price changes reflect conditions in a specific market as opposed to macroeconomic conditions. It is possible that prices can respond strongly and quickly to sector-specific shocks while adjusting slowly and imperfectly to aggregate shocks.[3] Boivin, Giannoni, and Mihov (2009) find evidence in support of this idea. They study disaggregated data on consumer and producer prices and examine how consumer prices respond to common and sector-specific shocks. They find that most price changes are driven by sector-specific developments rather than aggregate conditions and disaggregated prices are sticky in response to macroeconomic conditions.

18.3 The New Keynesian model

Early Keynesian theories were based on simple relationships between aggregate variables that did not have a close connection to the microeconomic decisions made by individuals.

[3]This may occur due to information frictions, e.g. sellers are unaware of all the macroeconomic developments that may impact the optimal price. Price adjustments may also be imperfect if firms want to keep their prices near those of other firms who are not simultaneously updating their prices. This force is known as "real rigidities" and it amplifies the effect of nominal rigidities.

Research in the 1980s and 1990s developed the **New Keynesian** model that is based on the same principles of optimization and equilibrium that underlie modern macroeconomics and incorporate frictions that interfere with the immediate and perfect adjustment of prices.

Nominal rigidities can be modeled in several ways. One natural way to model them is to assume prices are fixed for a certain period of time. For example, many employment contracts are reviewed and adjusted at an annual frequency (see Grigsby, Hurst, and Yildirmaz, 2021). Apartment rents are another type of transaction with long-term contracts. The same may be the case for wholesale prices where prices of goods are subject to longer term contracting between firms. Taylor's (1980) overlapping-contracts model adopts such a view of nominal rigidities and assumes that each price is fixed for some number n periods and therefore each period $1/n$ of the prices in the economy is updated. This assumption requires the modeler to keep track of n different prices. A simpler approach is to assume that a fraction $1-\theta \in [0,1]$ of the prices in the economy is updated each period, but unlike Taylor's model, the prices that are updated each period are randomly drawn from the existing prices. This assumption, which was introduced by Calvo (1983), leads to a much more tractable model because it is no longer necessary to keep track of the previously set prices as we will explain below.

Environment. A representative household has preferences for consumption and leisure as represented by

$$U_0 = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi} \right], \tag{18.1}$$

where C_t denotes consumption and L_t is labor supply and $\beta \in [0,1]$. The household chooses how much to work, how much to save, and how much to consume. The vehicle for savings is a one-period nominal bond that pays interest i_t in period $t+1$ and is in zero net supply.

The household consumes a final good, which is produced by a representative competitive firm out of a continuum of intermediate inputs indexed by $j \in [0,1]$. The final goods producer operates a constant elasticity of substitution production function. Letting Y_t be the quantity of final goods produced, the production function is given as:

$$Y_t = \left(\int_0^1 y_{j,t}^{\frac{\varepsilon-1}{\varepsilon}} dj \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{18.2}$$

where $y_{j,t}$ is the quantity of the jth intermediate input used and $\varepsilon > 1$ is the elasticity of substitution between varieties. The intermediate inputs are produced according to the production function

$$y_{j,t} = A_t \ell_{j,t}, \tag{18.3}$$

where $\ell_{j,t}$ is the amount of labor used in producing good j and A_t is an aggregate productivity that follows a first-order Markov process. The real marginal cost of producing one unit of intermediate goods is common across firms and equal to w_t/A_t, where w_t is the real wage denominated in units of final goods.

To model nominal rigidities we necessarily have to depart from perfect competition. After all, if all firms are price takers, no firm can be said to set its price, and when prices are sticky, firms need to stand ready to satisfy demand at the possible preset price which a competitive

firms may not be willing or able to do. We therefore adopt monopolistic competition as a convenient framework in which to model market power. We have seen the production function (18.2) before in Chapter 6. Using the argument that we introduce there, the final goods producer's cost-minimization problem yields a price index

$$P_t = \left(\int_0^1 P_{j,t}^{1-\varepsilon} dj \right)^{1/(1-\varepsilon)},$$

where P_t is the cost of producing one unit of the final good and $P_{j,t}$ is the price of a unit of the jth intermediate good.[4] As the final goods producer is competitive, it sells its output at marginal cost, and P_t is therefore also the price of a unit of final goods.

It also follows from the final goods producer's cost-minimization problem that the demand for intermediate good j is given by the following iso-elastic function of the good's relative price:

$$y_{j,t} = \left(\frac{P_{j,t}}{P_t} \right)^{-\varepsilon} Y_t. \tag{18.4}$$

Note that the price elasticity of the demand facing an intermediate goods producer is ε.

Both $P_{j,t}$ and P_t are nominal prices meaning they are set in terms of units of money. Money does not feature in this version of the New Keynesian model except as the unit of account for prices and bonds. Here we are making use of the cashless-limit argument introduced in Chapter 16, in which case households hold zero cash balances yet money can serve as the numeraire. However, extending the model with money demand derived from a money-in-the-utility-function assumption would produce exactly the same equilibrium as long as the marginal rate of substitution between consumption and leisure is independent of real cash balances.[5]

The key assumption that distinguishes New Keynesian models is that they feature nominal rigidities—that is, one or more prices in the economy is sticky. These nominal rigidities can apply to goods prices, to wages, or both; but here we will assume that the prices of intermediate inputs are sticky with the price-setting friction modeled as in Calvo (1983). Specifically, an intermediate goods producer has an i.i.d. probability $1 - \theta \in [0, 1]$ of being able to update its price each period. In the monopolistic competition model, the intermediate goods producers who are allowed to adjust prices in a given period then set prices taking as given the actions of other firms, input prices, and all aggregate variables. Firms serve whatever demand they face given the price they are quoting. Thus, in the short run, output is "demand determined" in the New Keynesian model.

The government is the final participant in this economy. The government sets the nominal interest rate and its choices are often represented through an interest rate rule that specifies the interest rate as an explicit function of macroeconomic variables such as the rate of inflation. We defer making detailed assumptions about the monetary policy rule.

[4] Formally, P_t is the derivative of the cost function for the final goods producers with respect to the quantity produced.

[5] Adding money demand would simply determine money supply.

Decision problems. The household's decision problem is to maximize

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{C_t^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi} \right],$$

subject to the budget constraint

$$B_{t+1} + P_t C_t = P_t w_t L_t + (1 + i_{t-1}) B_t + P_t \int m_{jt} dj \qquad \forall t,$$

where B_{t+1} is the nominal bonds held from date t to date $t+1$ and m_{jt} is the date-t profit earned by firm j. In addition to the budget constraint, the household must also satisfy a no Ponzi game condition of the kind introduced in Chapter 4. It is convenient to rewrite the budget constraint in real terms by dividing both sides by P_t

$$b_{t+1} + C_t = w_t L_t + \frac{1 + i_{t-1}}{1 + \pi_t} b_t + \int m_{jt} dj,$$

where $a_t \equiv B_t / P_{t-1}$ is the real value of savings in period $t-1$ and $1 + \pi_{t+1} \equiv P_{t+1}/P_t$ is the gross rate of inflation between periods t and $t+1$.[6]

This decision problem results in two optimality conditions. The Euler equation

$$C_t^{-\sigma} = \beta \mathbb{E}_t \left[\frac{1 + i_t}{1 + \pi_{t+1}} C_{t+1}^{-\sigma} \right] \tag{18.5}$$

and the intratemporal labor supply condition

$$C_t^{-\sigma} w_t = L_t^{\psi}. \tag{18.6}$$

The Euler equation in (18.5) implies that the intertemporal path of consumption is determined by the (gross) real interest rate, $(1 + i_t)/(1 + \pi_{t+1})$. Hence, household's intertemporal choices are no different than in standard flexible price models encountered earlier in this book. Likewise, the labor supply condition in equation (18.6) is also standard and equalizes households' marginal rate of substitution between consumption and work with the real wage. Nominal rigidities affect these conditions only through equilibrium changes in real wages and real interest rates.

An intermediate goods firm chooses how to price its good when it has the opportunity to reset its price. As usual, it is only relative prices that matter so we will define $p_{jt} \equiv P_{jt}/P_t$. Suppose a firm has a relative price p_{jt} at date t. If at date $t+1$ the firm does not update its nominal price, its new relative price will be

$$p_{jt+1} = \frac{P_{jt}}{P_{t+1}} = \frac{P_{jt}}{P_t} \frac{P_t}{P_{t+1}} = \frac{p_{jt}}{1 + \pi_{t+1}}.$$

If the firm instead does update its price at date $t+1$, it faces the same decision problem as any other firm that is updating its price that period because there are no firm-specific

[6] Note that in the standard notation used here, i_t applies to savings from t to $t+1$ while π_t refers to the inflation between $t-1$ and t. Both variables are determined at date t.

state variables other than the price they set in the past, which is no longer relevant to the firm updating its price. Let P_{t+1}^R be the nominal price set by firms who update their prices in $t+1$, which is known as the **reset price.** Let $p_{t+1}^R = P_{t+1}^R/P_{t+1}$ be the corresponding relative price.

In period t, the real profits of firm j are

$$m_{jt} = p_{jt}y_{jt} - w_t\ell_{jt} = (p_{jt} - w_t/A_t)y_{jt} = (p_{jt} - w_t/A_t)p_{jt}^{-\varepsilon}Y_t,$$

where the we have used eq. (18.3) to substitute for ℓ_{jt} and eq. (18.4) to substitute for y_{jt}.

The value function of a firm with relative price p is

$$V(p,\mathcal{S}) = u'(C(\mathcal{S}))\left(p - \frac{w(\mathcal{S})}{A(\mathcal{S})}\right)p^{-\varepsilon}Y(\mathcal{S}) + \beta\mathbb{E}\left[\theta V\left(\frac{p}{1+\pi(\mathcal{S}')},\mathcal{S}'\right) + (1-\theta)V\left(p^R(\mathcal{S}'),\mathcal{S}'\right)\right],$$

where $\mathcal{S}$ is the aggregate state, which evolves independently of an individual firm's decisions because of the monopolistic competition assumption. On the right-hand side of this Bellman equation, the first term is the period payoff, which is the profits earned during the period valued at the marginal utility of consumption of the representative household (the owner of the firm). The second term is the continuation value, which has two components. The first component represents the value if the firm does not update its price next period while the second component is the continuation value if it does update.

A firm that updates its price maximizes $V(p,\mathcal{S})$ therefore finds the price that solves:

$$p^R(\mathcal{S}) = \arg\max_p V(p,\mathcal{S}).$$

As shown in the appendix, the first-order necessary condition for the choice of the optimal reset price can be expressed as:

$$0 = \mathbb{E}_t\sum_{\tau=t}^{\infty}(\beta\theta)^{\tau-t}u'(C_\tau)Y_\tau\left[{p_{t,\tau}^R}^{-\varepsilon} - \varepsilon{p_{t,\tau}^R}^{-\varepsilon-1}\left(p_{t,\tau}^R - \frac{w_\tau}{A_\tau}\right)\right],$$

where $p_{t,\tau}^R \equiv P_t^R/P_\tau = p_t^R/\prod_{s=t+1}^{\tau}(1+\pi_s)$, is the relative price at date τ of a firm that last updated its price at date t. This condition is interesting. First, the term in square brackets times Y_τ is the marginal change in profit from a change in $p_{t,\tau}^R$. In a flexible price model, this term would equal zero period by period. But, because prices are sticky, firms aim at hitting this condition "on average" which means weighting profits in future periods by the probability $\theta^{\tau-t}$ that price will remain in effect that far in the future (and discounting future marginal profit contributions by the owners' marginal rate of intertemporal substitution). Secondly, firms are more forward looking when prices are stickier. This result is intuitive. In a flexible price setting, firms simply solve static optimization problems that result in equalizing marginal revenue and marginal costs period by period. With sticky prices, they now need to consider how their current choice of price may affect future profits. For example, firms will adjust prices today in response to information about marginal costs in the future.

Rearranging, we obtain

$$p_t^R\mathbb{E}_t\sum_{\tau=t}^{\infty}(\beta\theta)^{\tau-t}u'(C_\tau)Y_\tau\left(\frac{P_\tau}{P_t}\right)^{\varepsilon} = \frac{\varepsilon}{\varepsilon-1}\mathbb{E}_t\sum_{\tau=t}^{\infty}(\beta\theta)^{\tau-t}u'(C_\tau)Y_\tau\left(\frac{P_\tau}{P_t}\right)^{1+\varepsilon}\frac{w_\tau}{A_\tau}.$$

We can then express the condition for the optimal reset price as:

$$p_t^R = \frac{p_t^N}{p_t^D} \tag{18.7}$$

where:

$$p_t^N = \frac{\varepsilon}{\varepsilon - 1} \frac{w_t}{A_t} u'(C_t) Y_t + \theta\beta\mathbb{E}_t(1+\pi_{t+1})^{1+\varepsilon} p_{t+1}^N$$

and

$$p_t^D = u'(C_t) Y_t + \theta\beta\mathbb{E}_t(1+\pi_{t+1})^{\varepsilon} p_{t+1}^D.$$

Market clearing. There are three markets at each date: the labor market, the goods market, and the bond market. Labor market clearing requires that the total labor supplied by households equals the total labor used in production

$$L_t = \int \ell_{j,t} dj.$$

As consumption is the only use for final goods, goods market clearing requires that production equals total consumption

$$Y_t = C_t. \tag{18.8}$$

Finally, bond market clearing requires that the net demand for bonds is zero.

Aggregation. In the Calvo model, firms are given the opportunity to re-optimize their prices with probability $1-\theta$ and it is assumed that the process for the arrival rate of this opportunity is memoryless (ie. it follows a Poisson process). Hence, every period, firms are split randomly into a group of firms that cannot adjust their prices and another group that can. Let $\mathcal{J}_t$ be the set of firms that update their prices in period t. It follows then that the price level can be expressed as:

$$P_t^{1-\varepsilon} = \int_j P_{j,t}^{1-\varepsilon} dj = \underbrace{\int_{j \notin \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj}_{\text{non-adjusters}} + \underbrace{\int_{j \in \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj}_{\text{adjusters}}$$

where each of these two terms are given as:

$$\int_{j \notin \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj = \theta P_{t-1}^{1-\varepsilon}$$

and

$$\int_{j \in \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj = (1-\theta)\left(P_t^R\right)^{1-\varepsilon}.$$

Here we have used the assumption that firms are chosen randomly to adjust their prices to derive the expression for $\int_{j \notin \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj$. The expression for $\int_{j \in \mathcal{J}_t} P_{j,t}^{1-\varepsilon} dj$ exploits the fact that all firms that optimize their prices at the same date, choose the same price. It follows from this that:

$$P_t = \left[\theta P_{t-1}^{1-\varepsilon} + (1-\theta)\left(P_t^R\right)^{1-\varepsilon}\right]^{1/(1-\varepsilon)}. \tag{18.9}$$

This expression makes it clear that the price level is sticky since it displays inertia. The higher is the probability that the firm cannot adjust prices in any given period, θ, the larger is the backward looking component of the price level. Yet, despite this, inflation is purely forward looking in this model. To see this, we can also exploit the above result to derive a relationship between inflation and the relative reset price:

$$p_t^R = \frac{\left[\frac{P_t^{1-\varepsilon} - \theta P_{t-1}^{1-\varepsilon}}{1-\theta}\right]^{1/(1-\varepsilon)}}{P_t} = \left[\frac{1-\theta(1+\pi_t)^{\varepsilon-1}}{1-\theta}\right]^{1/(1-\varepsilon)}$$

which we can rearrange to give:

$$1+\pi_t = \left[\frac{1-(1-\theta)\left(p_t^R\right)^{1-\varepsilon}}{\theta}\right]^{1/(\varepsilon-1)}. \tag{18.10}$$

Since firms are purely forward looking when setting p_t^R, inflation is purely forward looking, too. As a result, we do not need to keep any record of past prices (or inflation) when determining the current (or future) inflation rates.

Since only a fraction of firms can adjust prices, there can be dispersion of prices across firms. This price dispersion across intermediate input producers is a source of inefficiency as the final goods producers will over-utilize those inputs that have low relative prices and under-utilize those with high relative prices. To see this, integrate equation (18.3) across j

$$\int y_{j,t} dj = A_t \int \ell_{j,t} dj = A_t L_t,$$

where the second equality uses the labor market clearing condition. Now substitute (18.4) for $y_{j,t}$ to arrive at

$$Y_t = \frac{A_t}{D_t} L_t, \tag{18.11}$$

where

$$D_t \equiv \int \left(\frac{P_{j,t}}{P_t}\right)^{-\varepsilon} dj \geq 1$$

is a measure of price dispersion. In particular, if $P_{j,t}$ is common across all firms, as would happen in this model if intermediate goods prices were flexible, then there is no price dispersion $P_{j,t} = P_t$ for all j and $D_t = 1$. In this case it follows that $Y_t = A_t L_t$. When there is price dispersion, however, D_t will be greater than one, which reduces the effective productivity of the economy. This loss of efficiency is due to the concave production function for final goods and the misallocation of labor across intermediate varieties. By the logic that the updated prices are randomly selected we have

$$D_t = \int_{j\in\mathcal{J}_t} p_t^{R-\varepsilon} dj + \int_{j\notin\mathcal{J}_t} \left(\frac{P_{j,t-1}}{P_{t-1}}\frac{P_{t-1}}{P_t}\right)^{-\varepsilon} dj = (1-\theta)p_t^{R-\varepsilon} + \theta(1+\pi_t)^{\varepsilon} D_{t-1}. \tag{18.12}$$

State variables. There are two state variables of the aggregate economy: the current level of TFP, A_t, and the degree of price dispersion D_{t-1}. These states evolve according to the exogenous law of motion for A_t and equation (18.12), respectively.

Equilibrium. An equilibrium of this economy involves stochastic processes for C_t, L_t, Y_t, w_t, π_t, i_t, A_t, D_t that satisfy (18.5), (18.6), (18.8), (18.10), (18.11), (18.12), the monetary policy rule, and the exogenous law of motion for A_t; in addition, an equilibrium requires a policy rule p^R and a value function V that solve the price-setting problem.

Flexible-price equilibrium and output gap. In order to complete the specification of the environment we need a monetary policy rule. An important benchmark for policy in this context is the counterfactual outcome with fully flexible prices. We introduce that now as common monetary policy rules make reference to this benchmark.

The flexible-price economy is a special case where $\theta = 0$ so all prices are updated every period. When an intermediate goods firm sets its price in period t, it knows it will for certain be able to adjust its price again in $t+1$ and the only consideration is maximizing profits in period t. This static profit maximization problem results in setting a constant markup $\mu \equiv \varepsilon/(\varepsilon - 1)$ over marginal cost as we saw in Chapter 6.[7] So $p_t^R = \mu w_t / A_t$ and the markup is positive as long as $\varepsilon > 1$ which we have imposed earlier. As all firms face the same profit maximization problem when prices are flexible, they all choose to set the same price and the definition of the price index reduces to $P_t^n = P_t^{R,n}$ or $p_t^{R,n} = 1$ where we use the superscript "n" to indicate variables in the flexible price, or "natural," equilibrium. It then follows that $w_t^n / A_t = \mu^{-1}$. Moreover, as all firms set the same price, there is no efficiency loss from price dispersion. We can then solve the following system of equations to obtain the equilibrium:

$$\begin{aligned} w_t^n &= A_t/\mu \\ Y_t^n &= C_t^n \\ Y_t^n &= A_t L_t^n \\ C_t^{n-\sigma} w_t^n &= (L_t^n)^{\psi}, \end{aligned}$$

where the second, third, and fourth equations are the aggregate resource constraint, the aggregate production function, and the intratemporal labor supply condition. The solution to this system is:

$$Y_t^n = \mu^{-1/(\sigma+\psi)} A_t^{(1+\psi)/(\sigma+\psi)}. \tag{18.13}$$

Let us also define the output gap as

$$x_t \equiv \log Y_t - \log Y_t^n \tag{18.14}$$

i.e. the log of level of output in the sticky-price equilibrium relative to the natural level.

[7] You can verify this by solving $\max_p \{(p - w_t/A_t)\, p^{-\varepsilon} Y_t\}$.

Interest rate rules. An interest rate rule describes how monetary policy is set. These can take several forms, but the simplest version is to impose a relationship between nominal interest rates and other variables in the model. For example,

$$i_t = \bar{i} + \phi_\pi(\pi_t - \pi^*) + \phi_x x_t,$$

where π^* is the inflation target and $\bar{i} = \beta^{-1}(1 + \pi^*) - 1$ is the nominal interest rate that would satisfy the Euler equation (18.5) in a steady state with constant inflation π^*. The coefficients ϕ_π and ϕ_x determine how strongly nominal interest rates react to deviations of inflation from its target and to the output gap. Interest rate rules are often called **Taylor rules** in light of Taylor's (1993) finding that such a rule with $\phi_\pi = 1.5$ and $\phi_x = 0.5$ provided a good description of the monetary policy choices of the Federal Reserve between 1987 and 1992.[8]

Monetary policy shocks. Here we are building the baseline New Keynesian model. More elaborate versions of this model are often used in central banks to analyze potential monetary policy strategies. Researchers and policymakers are therefore often interested in questions of the form "what would happen if we raise interest rates taking the economic conditions as given?" To answer this question, we want to calculate the consequences of a shift in interest rates that is not the result of a shift in the economic fundamentals that appear in the interest rate rule. The solution is to add an exogenous shock to the interest rate rule.

The three-equation model. The model we have presented above is often called the **three-equation model** because a first-order approximation of the model around a zero-inflation steady state is summarized by the following three equations. First, we have a log-linearized version of the consumption Euler equation

$$x_t = \mathbb{E}_t x_{t+1} - \frac{1}{\sigma}(\hat{i}_t - \mathbb{E}_t \pi_{t+1} - r_t^n), \tag{18.15}$$

where x_t is the output gap and

$$r_t^n = -\log\beta + \frac{\sigma(1+\psi)}{\sigma+\psi}(\hat{A}_t - \mathbb{E}_t \hat{A}_{t+1})$$

is the real natural rate of interest.[9] This equation represents the demand-side of the economy and is sometimes called the **IS curve** in a reference to older IS-LM models. Notice that

[8] The coefficients in Taylor's specification correspond to annual rates of inflation and annualized interest rates. Interest rates and inflation rates are measured in percentage points per unit of time while the output gap is simply measured in percentage points. Therefore, if you change the length of a unit of time, you need to adjust the coefficients accordingly. For example, in a quarterly model, the equivalent rule would be $\phi_\pi = 1.5$ and $\phi_x = 0.125$.

[9] To derive this equation, define $\hat{i}_t = \log(1 + i_t)$. Then take logs of both sides (18.5) noting that in a first-order perturbation solution we assume the shocks are arbitrarily small so we can take logs inside the expectation operator. Finally, we need the definition of the output gap from equations (18.13) and (18.14), which when linearized give $x_t = \hat{Y}_t - \frac{1+\psi}{\sigma+\psi}\hat{A}_t$. Using this to substitute for $\hat{Y}_t$ yields the equation above. Note that r_t^n is equal to the real interest rate that solves the linearized Euler equation when consumption takes its natural level.

output at date t is forward-looking (it depends on $\mathbb{E}_t \hat{Y}_{t+1}$) and is decreasing in the real interest rate where $1/\sigma$ is the elasticity of intertemporal substitution.

Next, as we show in the appendix, we use the first-order condition of the firm's price-setting problem and other equilibrium conditions of the model to arrive at a New Keynesian Phillips curve

$$\pi_t = \kappa x_t + \beta \mathbb{E}_t \left[\pi_{t+1}\right], \tag{18.16}$$

where the composite parameter $\kappa \equiv (1-\theta)(1-\beta\theta)(\sigma+\psi)/\theta$ is usually referred to as the slope of the New Keynesian Phillips curve. Here we see that inflation is forward-looking and increasing in the current output gap. The slope of the Phillips curve, κ, is larger if prices are more flexible (lower θ) or if marginal cost is more sensitive to the level of production.

Lastly, we have an interest rate rule. For our simulations here we will assume

$$\hat{i}_t = \phi_\pi(\pi_t - \pi^*) + \phi_x x_t + \omega_t, \tag{18.17}$$

where ω_t is an AR(1) monetary policy shock. In addition to these three core equations, we also have exogenous processes for TFP and the monetary policy shock. The three-equation representation of the model, (18.15)-(18.17), is often used as a small-scale model of the economy to explore qualitative features of business cycles and monetary policy.

Determinacy. A key consideration for any monetary policy rule is whether it yields a unique equilibrium. Suppose that, for some reason, inflation expectations are elevated despite expectations of a zero output gap. Could that lead to high inflation today? It all depends on the behavior of the nominal interest rate. If the nominal interest rate is unresponsive, high inflation expectations translate to low real interest rates, which through the IS curve lead to a positive output gap. The positive output gap then puts upward pressure on current inflation through the Phillips curve. By this logic, some small expectation of higher inflation in the far future could justify expectations of high inflation (and positive output gaps) all the way back to the present. In the model, nothing pins down the expectations of inflation in the infinitely far future so we indeed have multiple equilibria when nominal interest rates are unresponsive. In order to prevent this outcome, the interest rate rule must respond sufficiently strongly so that the real interest rate *rises* when inflation rises thereby stabilizing the economy. The three-equation model has a unique equilibrium if and only if

$$(1-\beta)\phi_x + \kappa(\phi_\pi - 1) > 0 \tag{18.18}$$

(see the appendix for a derivation). Note that $\phi_\pi > 1$ is sufficient for this condition to hold and necessary if $\phi_x = 0$. This condition is known as the **Taylor principle**.

The role of nominal demand in determining output. As we mentioned in the introduction to this chapter, when prices adjust imperfectly, changes in nominal variables can have real effects. Traditionally, this point is often discussed in terms of the money supply—if money is injected, a frictionless model would imply an immediate adjustment of the price level with no effect on any real variables. This outcome is often called the classical dichotomy between real and nominal sides of the economy. With nominal rigidities, this logic breaks

down as the increase in nominal demand from the increase in the money supply is not immediately undone by an increase in the price level. Here we will make a similar argument in terms of nominal interest rates. Suppose the government announces a lower path for nominal interest rates going forward (e.g. an expansionary monetary policy shock). As we saw in equation (18.5), the path of aggregate consumption (and therefore aggregate output) is fully determined by the path of real interest rates. Therefore, if the classical dichotomy were to hold, it would have to be the case that path of expected inflation would immediately fall so as to leave real interest rates unchanged. But if we solve equation (18.16) forward, we see inflation at each date reflects expectations of current and future output gaps, which would be zero under the classical dichotomy. In fact, the model predicts that with lower real interest rates there will be a positive output gap and an increase in inflation that can reinforce the decline in nominal interest rates absent a counter-veiling response of the monetary policy rule.

18.4 Monetary policy strategies

The New Keynesian model has an active role for government. After all, a key equation in the model is the monetary policy rule. We now describe what the role for policy is in the New Keynesian model and how this is reflected in some real-world monetary policy strategies.

18.4.1 Policy objectives

The equilibrium of the New Keynesian model can be inefficient for two reasons (i) the total level of production can deviate from the efficient level and (ii) labor can be misallocated across intermediate inputs. To see this in more detail, consider the planner's problem

$$\max_{C_t, Y_t, L_t} \frac{C_t^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi}$$

subject to

$$C_t = Y_t = A_t L_t.$$

This is a static problem because the planner only has to choose how to allocate labor each period and in the aggregate there is no way to move resources across time. The constraint on the planner uses the fact that when L_t units of labor are used to produce each intermediate good, the quantity of final goods that can be produced is $A_t L_t$. The first-order condition of this problem leads to

$$Y_t^* = A_t^{(1+\psi)/(\sigma+\psi)}, \tag{18.19}$$

where a star denotes the first-best allocation.

To focus on the possibility of an inefficient level of aggregate production, suppose there is no price dispersion so $Y_t = A_t L_t$. Then using (18.19), (18.6), and $C_t = Y_t$ we obtain

$$Y_t = \left(\frac{w_t}{A_t}\right)^{1/(\sigma+\psi)} Y_t^*.$$

In an efficient economy, the real wage would be equal to the marginal product of labor. The equation above shows that if the wage deviates from this efficient level, output will deviate from the efficient level. When the wage is too low, households will supply too little labor and too little will be produced. Comparing this to the natural level of output (18.13), we see that in the flexible price economy, output is too low because $w_t/A_t = 1/\mu < 1$. So monopoly power is one source of inefficiency. Notice that w_t/A_t is real marginal cost and if firms set prices equal to nominal marginal cost we would have $P_t = (P_t w_t)/A$ or $w_t/A_t = 1$.

Nominal rigidities affect the markups that firms charge and therefore the degree of inefficiency in the economy. To see this, suppose there is a decrease in nominal marginal cost. As some firms do not update their prices, their markups rise. In the aggregate, the increase in average markups is reflected in a lower w_t/A_t ratio and the economy is further from the efficient level of production. Over time, firms will update their prices and return their markups to the desired level μ and w_t/A_t will return to $1/\mu$.

Now consider the inefficiency due to price dispersion. Using the aggregate production function $Y_t = A_t L_t/D_t$, (18.19), (18.6), and $C_t = Y_t$ we obtain

$$Y_t = D_t^{-\psi/(\sigma+\psi)} \left(\frac{w_t}{A_t}\right)^{1/(\sigma+\psi)} Y_t^*.$$

This equation shows that there is an additional inefficiency, when $D_t \neq 1$. From the concavity of the final goods production function, the efficient use of labor is to produce equal amounts of all the intermediate inputs. The economy will deviate from this outcome if some intermediate goods producers have lower prices than others (see eq. 18.11) . The reason one producer may have a lower price than another is, for example, that one set its price more recently than the other and, in the intervening time, market conditions changed leading to a change in the optimal reset price. As the reset price fluctuates, so too will the inflation rate (see eq. 18.10). One goal for monetary policy then is to stabilize the inflation rate because this minimizes the efficiency loss from relative price dispersion.[10]

Define the *welfare-relevant output gap* as $x_t^w \equiv \log Y_t - \log Y_t^*$. Using equations (18.13) and (18.19) we see that

$$x_t^w = \log Y_t - \log Y_t^n + \log Y_t^n - \log Y_t^* = x_t - \frac{1}{\sigma+\psi}\log \mu.$$

The welfare-relevant output gap differs from the output gap because the natural level of output is distorted by monopoly power.

In the literature that followed Phillips' original contribution, Phillips (1958), it was perceived that there was a stable relationship between (wage) inflation and resource utilization (as measured by unemployment) which therefore presented a trade-off to policymakers. This notion was challenged by Friedman (1968) and Phelps (1967) who posited that the long-run Phillips curve is vertical (at the natural rate of unemployment) as real wages, which should be independent of inflation in the long-run, determine employment. If monetary policy attempts to persistently raise the level of output above the natural level, the results will be

[10]The Calvo model of nominal rigidities can generate large efficiency losses from price dispersion because a firm can be stuck with a very old price that is far out of alignment with the current price level. Models of menu costs in which firms can choose to change their prices subject to a cost tend to generate less price dispersion because firms with very misaligned prices will choose to change their prices.

high inflation and no actual increase in output. For these reasons, a common view is that monetary policy should focus on stabilizing the economy around the flexible-price level of activity even though this may not be the first best. One way of formalizing this view is to give the policymakers another policy tool that can address long-run inefficiencies while leaving monetary policy responsible for responding to aggregate shocks. In the New Keynesian model, a simple extension of the model is to suppose there is lump-sum tax on households that finances a constant production subsidy for intermediate goods producers. In particular, suppose that intermediate goods producers are given a subsidy $\tau^l = 1/\epsilon \in (0,1)$ so that their effective cost of labor is $(1-\tau^l)w_t/A_t$. The production subsidy induces them to produce more and, in the steady-state, the subsidy undoes the distortion from monopoly power.

In summary, in the rest of this chapter we will take the goals for policy to be to (i) bring the aggregate level of production to the natural level and to (ii) minimize the efficiency loss from price dispersion by stabilizing inflation. Locally around a zero-inflation steady state, the welfare of the representative household in the basic New Keynesian model can be expressed as[11]

$$U \approx -\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\pi_t^2 + \frac{\kappa}{\varepsilon} x_t^2 \right]. \tag{18.20}$$

Welfare decreases with inflation or deflation because it results in price dispersion. Welfare also decreases with a positive or negative output gap because the level of production differs from the optimal level.

18.4.2 The divine coincidence

In the baseline New Keynesian model, it is possible to have zero inflation and zero output gaps at all times. The New Keynesian Phillips curve, (18.16), is the key to this argument. From that equation we can see that if there is no output gap at any date, then there will be no inflation today or in any future period. This aspect of the model—that there is no trade off between output gap and inflation stabilization—is known as the **divine coincidence**.

The standard interpretation of the divine coincidence is that the model is somewhat too simple and abstracts from the features of the economy that lead to a meaningful trade-off between inflation stabilization and output stabilization. What might those features be? To break the divine coincidence we need a *time-varying* wedge between the flexible-price level of output and the efficient level of output. Productivity shocks themselves do not create such a wedge because they affect both the flexible-price level of output as well as the efficient level of output in equal proportions. We say the wedge needs to be time-varying because a constant distortion requires a permanent change in the level of activity and is therefore not an issue for which monetary policy is well suited.

An extension of the model, which is often discussed due to its simplicity, involves shocks to the elasticity of substitution between varieties of intermediate goods. The resulting time-varying market power affects Y_t^n but does not change Y_t^* leading to a time-varying gap between the socially efficient level of production and the flexible-price level. In the linearized

[11] Here we use a second order approximation around a steady state in which the monopoly distortion has been corrected through the labor subsidy $\tau^l = 1/\epsilon$. See Woodford (2003b).

version of the model, these shocks appear as a "cost-push" shock—an exogenous term that is appended to the Phillips curve

$$\pi_t = \kappa x_t + \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \eta_t \tag{18.21}$$

(see Steinsson, 2003, for a derivation). With this added term, it is no longer possible to stabilize inflation and output perfectly because a policy of setting $x_t = 0$ at all dates no longer leads to $\pi_t = 0$. The divine coincidence also breaks if the flexible-price economy does not respond to shocks in the efficient way; for example because there are frictions in the determination of wages (see Blanchard and Galí, 2007).

18.4.3 Inflation targets and price level targets

Around the world, most central banks follow a version of a monetary policy strategy called **flexible inflation targeting**. This strategy can be summarized as minimizing deviations of inflation from a target level while also minimizing deviations of output from its natural level. These goals can be formalized in the objective function

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[(\pi_t - \pi^*)^2 + \lambda x_t^2\right], \tag{18.22}$$

where π^* is the inflation target and λ is a parameter that determines the relative weight placed on the two goals. The policymaker then seeks to minimize this loss function, which closely resembles eq. (18.20).

When inflation is unpredictable, it is risky to agree to a long-term nominal contract because the real values stipulated by the contract are unpredictable. This rationale would argue that it is more valuable for monetary policy to minimize the uncertainty over the price level than to target an inflation rate. Under price-level targeting, the central bank seeks to return the price level to a specific path. For example, consider the interest rate rule

$$i_t = \bar{i} + \phi_P(P_t - P_t^*) + \phi_x x_t$$

where P_t^* is the price-level target, which could be the price level in some base year scaled by a constant annual inflation rate. This rule dictates that the central bank raises interest rates whenever the price level is above target, which would, all else equal, put downward pressure on inflation and move the economy back toward the target. In addition to providing more certainty about the price level, price-level targeting is appealing because it makes clear that future policy will undo unwanted changes in the price level—something that is also useful under inflation targeting as we will see next.

18.4.4 Expectations, commitment, and time consistency

In the New Keynesian model, the private sector is forward looking: current inflation depends on expectations of future inflation and current demand depends on expectations of future real interest rates. Therefore, expectations of what monetary policymakers will do in the future affect the macroeconomic outcomes today. If the private sector expects interest rates to be high in the future *ceteris paribus*, output and inflation will be lower today.

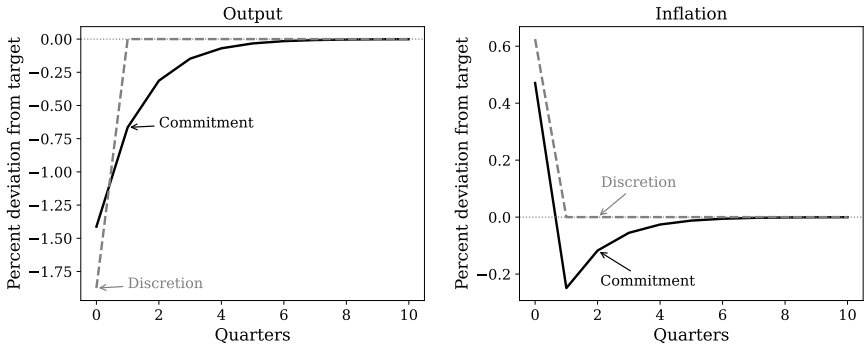


Figure 18.1: Response of output and inflation to a transitory cost push shock under commitment and discretionary policies.

Notes: These paths are simulated using the parameters $\beta = 0.99$, $\kappa = 0.2$, $\varepsilon = 3$.

Suppose a central bank pursues the inflation targeting objective (18.22) and the economy is hit by a cost-push shock at date 0. That is, we consider the Phillips curve (18.21) with $\eta_t > 0$ for just one period and zero thereafter. Now consider two possible monetary policy responses. In the first, the central bank raises interest rates at date 0 and then returns them to their long-run level. We will call this the "discretionary" policy for reasons that will become clear. In the second case, the central bank raises interest rates by less, but only gradually normalizes them. We will call this the commitment policy. Figure 18.1 plots the two cases. Under the discretionary policy, inflation rises at date 0 and then returns to zero. The linearized version of the model with the discretionary policy has no endogenous state variables so once the cost-push shock dissipates, the economy immediately returns to steady state.[12] The Phillips curve at date 0 is

$$\pi_0 = \kappa x_0 + \eta_0,$$

where we have used $\pi_1 = 0$ as the economy returns to steady state in the next period. In this case, the policymaker can only use x_0 to lean against the cost-push shock. As we see in the figure, output is reduced so as to dampen the inflationary effects of the shock.

Under the commitment policy, inflation is positive at date 0 and then negative for several future periods. The Phillips curve at date 0 is

$$\pi_0 = \kappa x_0 + \beta \pi_1 + \eta_0,$$

where now π_1 appears. Under this policy, the central bank reduces π_1 in order to reduce π_0 with less impact on x_0. As we see in the figure, there is less inflation at date 0 but also a smaller decline in output at date 0. In order to reduce π_1, the central bank keeps interest rates persistently higher than normal so that output and inflation are persistently lower.

While the outcomes from the commitment policy are better than those from the discretionary policy at date 0, they are worse at future dates. Under the discretionary policy, there are no deviations in output or inflation at any date $t \geq 1$ while, under the commitment policy, output and inflation are too low relative to the targets. This brings us to a time-consistency problem. At date 0, the central bank would like to announce the commitment

[12] In the non-linear model, the degree of price dispersion is an endogenous state variable.

policy, but it would like to switch to the discretionary policy at date 1. If the private sector anticipates this switch in policy, then the benefits of the commitment policy at date 0 are unattainable because the central bank cannot convince the private sector that π_1 will be negative. The central bank can only achieve the better outcomes if it is able to commit at date 0 to have tight monetary policy at future dates even though it will not want to do that when those future dates arrive.

Finally, under the discretionary policy, the price-level jumps at date 0 and then remains constant thereafter. Under the commitment policy, the price level rises, but then falls subsequently as inflation is negative for several periods. In fact, if you accumulate the inflation rates in the figure, you find that the long-run price level is unaffected. This result demonstrates that long-run price stability is useful for a central bank even if it is pursuing an inflation targeting framework because it serves to stabilize inflation expectations and therefore helps stabilize inflation rates.

18.5 Aggregate evidence of nominal rigidity

We now review the aggregate evidence on price rigidity and the response of real variables to nominal shocks in the form of changes in the nominal interest rate targeted by monetary policy. As we will see, the baseline model is qualitatively consistent with this evidence—in contrast to a flexible-price model—but also has shortcomings.

18.5.1 The macroeconomic effects of monetary policy shocks

One of the most studied issues in macroeconomics is the dynamic effects of monetary policy shocks, which lead to exogenous changes in nominal interest rates. This topic attracts so much interest because it speaks to the relevance of the nominal rigidities that underlie the Keynesian perspective. In most models with flexible prices, nominal variables including nominal interest rates have no bearing on real outcomes (e.g. see equation (18.13)). Thus, the impact of changes in nominal interest rates on real variables provides information about the importance of nominal rigidities. Moreover, monetary policy plays a central role in modern macroeconomic policy and understanding the consequences of a change in interest rates is a crucial ingredient to real-world policy decisions.

The main challenge with assessing the effects of a change in interest rates is that monetary policy changes *in response* to developments in the economy. If we simply look at the correlations between variables, we will tend to find that higher nominal interest rates are associated with higher levels of inflation, but this may simply reflect endogeneity of monetary policy decisions (because central banks tend to increase interest rates when inflation rises). We are instead interested in the *causal* effect of interest rates on the economy. That is, we ask what would happen if interest rates increased for reasons unrelated to state of the economy? To answer this question, we need to empirically identify monetary policy shocks as opposed to systematic movements in interest rates in response to economic conditions.

Researchers typically identify monetary policy shocks using information that allows them to estimate the systematic or predictable component of interest rates. Removing this systematic component from the actual change in interest rates yields a residual that is interpreted

as a monetary policy shock. One strand of literature, which was pioneered by Kuttner (2001) and Gürkaynak, Sack, and Swanson (2005), is premised on (a) the fact that monetary policy decisions are announced at particular times known to researchers and (b) in a narrow time window around the announcement, changes in interest rates will be dominated by monetary policy as opposed to other economic news. Using financial data, we can make a forecast of the monetary policy decision just minutes before it is announced. This forecast reflects market expectations of how monetary policy will be conducted given current economic conditions—i.e. it is the market's view of the systematic policy response. If the announced decision differs from the forecast, it is due to a deviation of monetary policy from it's usual practice (as judged by financial markets).

Another strand of literature, starting with Romer and Romer (2004), uses central bank forecasts of inflation and other variables to estimate the endogenous component of policy. The motivation for this approach is that central banks often set policy based on their assessment of the economic outlook as reflected in economic forecasts. For example, if the forecast for inflation is elevated or unemployment is expected to be low, policymakers will raise interest rates. By regressing interest rate changes on the central bank forecasts, this approach estimates the systematic component of policy and the residuals from this regression can be interpreted as movements in interest rates that are not due to changes in the economic outlook.

Figure 18.2 plots impulse responses for output, inflation, and nominal interest rates following a monetary policy shock identified using the Romer and Romer method.[13] The left panel of the figure shows that contractionary monetary policy shocks lead to persistently higher nominal interest rates. The center panel of the figure shows that aggregate output declines. The right panel shows that inflation declines so real interest rates rise more than nominal interest rates.[14] So we find that a nominal shock affects real variables.

The basic New Keynesian model presented above is qualitatively consistent with the patterns revealed by the empirical estimates: higher nominal interest rates generate increases in real interest rates and reductions in aggregate demand and therefore a decline in output. With lower output, there is lower resource utilization leading to lower marginal costs and a decline in inflation. The timing of the responses, however, is quite different. Figure 18.2 shows model-simulated responses of interest rates, output, and inflation to a monetary policy shock (for now, just focus on the basic model and we will return to the sticky-wage model below). We chose the magnitude of the simulated shock and the monetary policy rule to roughly match the path of nominal interest rates that we estimated. In addition, we chose the other parameters of the model to roughly match the magnitudes of the responses of output and inflation.[15] As shown in the figure, the model-generated responses of output

[13] We use the implementation of this method by Wieland and Yang (2016) and regress the outcome of interest on the estimated monetary policy shocks and the lags of macroeconomic variables.

[14] The increase in inflation at very short horizons is known as the "price puzzle" and is a fairly common feature of empirical estimates of the effects of monetary policy shocks. It likely reflects reverse causation from inflation to interest rates that is not removed by the identification strategy. In some specifications, adding further control variables eliminates the price puzzle (see Sims, 1992).

[15] Here we use equations (18.15) and (18.16) and a policy rule $\hat{i}_t = \gamma \hat{i}_{t-1} + (1-\gamma)(\phi_\pi \pi_t + \phi_x x_t) + \omega_t$ and an AR(1) specification for ω_t to mimic the estimated path of interest rates. We use parameter values $\gamma = 0.65$, $\phi_\pi = 1.5$, $\phi_x = 0.125$, $\sigma = 5$, $\psi = 1$, $\theta = 1 - \frac{1}{8}$, $\beta = 0.99$, and an autocorrelation of 0.5 for ω_t.

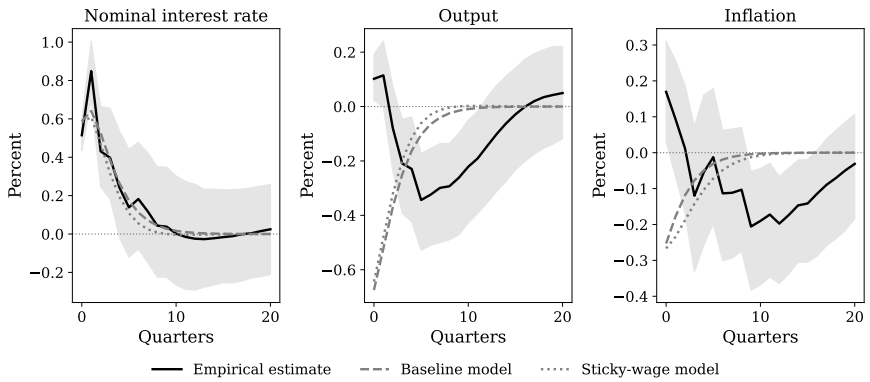


Figure 18.2: Empirical and simulated responses to a monetary policy shock.

Notes: The shaded areas around the empirical estimates are 90% confidence bands.

and inflation immediately fall in response to the monetary shock and then gradually return to steady state whereas the empirical responses feature delayed responses of output and inflation.

18.5.2 Price rigidity in the aggregate

At the start of this chapter we discussed evidence on the frequency of price changes in micro data. But does this micro-level rigidity translate to rigidity in the aggregate?

One way to answer this question is to rely on a general equilibrium model as discussed above and ask which values of structural parameters allow one best to "match" the data? One might, for example, study the estimated impulse response functions in Figure 18.2 and estimate the slope of the Phillips curve by comparing the magnitude of the inflation response to the magnitude of the output response. One implementation of this idea was carried out by Christiano, Eichenbaum, and Evans (2005). These authors construct a rich New Keynesian model and fit its structural parameters to match the impulse response functions following a monetary policy shock. Their model includes many extensions of the basic New Keynesian framework some of which we describe in the next section. Overall, they find that the data indicate a value of θ that is consistent with an average contract length of 2.5 quarters which is in the range of values suggested by the literature on the frequency of price changes discussed earlier. Yet, at the same time, they also find that rigidities in wage setting are crucial to explain the dynamics of inflation following a monetary policy shock.

An alternative approach is to estimate the New Keynesian Phillips curve directly in isolation from other parts of the model. Here, the question is how strongly changes in the output gap (or marginal costs of production) translate to changes in prices. Consider the Phillips curve

$$\pi_t = \kappa x_t + \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \eta_t$$

where we have included the shock η_t as we would in general not expect the Phillips curve to hold exactly as an empirical relationship. Our interest is in estimating $\kappa = (1-\theta)(1-\theta\beta)(\sigma+\psi)/\theta$ as this speaks to the strength of nominal rigidities.

The direct estimation of the Phillips curve requires one to address several challenges. Any measure of the output gap is bound to be associated with measurement error which induces an attenuation bias in our estimate of κ if we were simply to apply ordinary least squares. However, if we have an instrumental variable that is correlated with the true output gap and uncorrelated with the measurement error we can sidestep this source of bias. Inflation expectations are not directly observable either. To measure inflation expectations, the typical practice is to use statistical techniques to construct a forecast of π_{t+1} on the basis of information that is available at date t. Finally, the presence of the cost-push shock complicates matters. Suppose there is an inflationary cost push shock, $\eta_t > 0$, and that the central bank responds to this by restraining demand and inducing a negative output gap. In this scenario, there is reverse causation from inflation to the output gap.

The literature has adopted a variety of methods to overcome these challenges and it remains an active area of research. Galí and Gertler (1999) is representative of the approach that has been followed by many papers in the literature. This approach rewrites the Phillips curve as

$$\pi_t = \kappa x_t + \beta \pi_{t+1} + \underbrace{\eta_t - \beta \left(\pi_{t+1} - \mathbb{E}_t \left[\pi_{t+1} \right] \right)}_{\equiv \zeta_t},$$

where we have replaced expected inflation with the realization of π_{t+1} and included the expectational error $\pi_{t+1} - \mathbb{E}_t[\pi_{t+1}]$ in the error term, which we now denote ζ_t. If we assume $\mathbb{E}_{t-1}[\eta_t] = 0$, any lagged variable known at $t-1$ that predicts x_t or π_{t+1} can be a valid instrument.[16] This (strong) assumption allows us to use realized future inflation in place of expected inflation thereby obviating the need for measurement of inflation expectations and also prevents reverse causation as the lagged instruments are not correlated with η_t.

To measure the output gap, Galí and Gertler make use of the micro-foundations of the New Keynesian model, which say it is real marginal cost that is relevant in price setting. They start by positing a Cobb-Douglas aggregate production function:

$$Y_t = A_t K_t^{\alpha_K} L_t^{\alpha_L}$$

where K_t is the input of capital and α_K and α_L are the output elasticities with respect to the two factor inputs. Assuming that the labor input is flexible while the capital stock is predetermined in period t, cost minimization (together with firms being price takers in the input markets) implies that real marginal costs are given as:

$$mc_t = \frac{w_t}{\partial Y_t / \partial L_t} = \frac{s_t^L}{\alpha_L},$$

where $s_t^L = (w_t L_t)/Y_t$ is the labor share of income and w_t is the real wage. Thus, up to a first-order approximation, the log of real marginal costs are given as the log of the labor share.

[16] Let z_{t-1} be the instrument. The orthogonality condition requires that this instrument be uncorrelated with ζ_t, which we can verify as follows

$$\operatorname{cov}(z_{t-1}, \zeta_t) = \mathbb{E}\left[z_{t-1} \zeta_t \right] = \mathbb{E}\left[z_{t-1} \left\{ \eta_t - \beta \left(\pi_{t+1} - \mathbb{E}_t \left[\pi_{t+1} \right] \right) \right\} \right] = \mathbb{E}\left[z_{t-1} \eta_t \right] + \beta \mathbb{E}\left[z_{t-1} \left(\pi_{t+1} - \mathbb{E}_t \left[\pi_{t+1} \right] \right) \right] = 0,$$

where $\mathbb{E}\left[z_{t-1} \eta_t \right] = \mathbb{E}\left[z_{t-1} \mathbb{E}_{t-1} \left[\eta_t \right] \right]$ by the law of iterated expectations and $\mathbb{E}_{t-1}\left[\eta_t \right] = 0$ by assumption. Similarly $\mathbb{E}\left[z_{t-1} \left(\pi_{t+1} - \mathbb{E}_t \left[\pi_{t+1} \right] \right) \right] = \mathbb{E}\left[z_{t-1} \left(E_t \left[\pi_{t+1} \right] - \mathbb{E}_t \left[\pi_{t+1} \right] \right) \right] = 0$ by the law of iterated expectations.

Galí and Gertler estimate the Phillips curve using the generalized method of moments applied to quarterly U.S. data from 1960 to 1997. Their estimates imply an estimate of the frequency of price adjustment, $1-\theta$, in the range of 0.085-0.171 per quarter, which indicates much longer average price contract length than the estimates from disaggregated prices discussed earlier.

The New Keynesian Phillips curve implies inflation is purely forward-looking. Before the development of the New Keynesian model, typical specifications of the Phillips curve instead included lagged inflation rather than inflation expectations. This so-called "accelerationist" Phillips curve implies a high degree of inflation persistence. Forward- and backward-looking Phillips curves have very different implications for how inflation can be controlled. In the forward-looking case, a credible central bank can immediately reduce inflation by committing to a low long-run inflation rate and zero output gaps. In the latter case, inflation can only be reduced by imposing negative output gaps on the economy. To explore the issue of inflation persistence, Galí and Gertler also considered a "hybrid" New Keynesian Phillips curve in which they assume that a certain fraction of price setters are purely backward looking and simply set prices by updating the past average reset price (of forward-looking firms) with the past inflation rate. In this extension, they estimate the fraction of firms that are backward looking as opposed to forward looking. Their estimates place much more weight on forward-looking behavior than on backward-looking behavior.

The work of Galí and Gertler (1999) has been influential because it establishes a link between inflation and marginal costs as consistent with the price setting condition embedded in the New Keynesian model. However, there remains considerable uncertainty surrounding the Phillips curve parameters as changes in the data series, the sample period, or the econometric specification can result in considerable changes in the parameter estimates, see (see Mavroeidis, Plagborg-Møller, and Stock, 2014). One issue of note is that the correlation between the labor share and inflation appears to have weakened over time. In Figure 18.3 we show the dynamic correlations between the labor share and leads and lags of the inflation rate for U.S. quarterly data, 1960:1-2019:4. We compute cross correlations for both the whole sample and for an early sample (ending in 1997) and a late sample (starting in 1998). Consistent with the results of Galí and Gertler (1999),[17] there is a significant positive relationship between the labor share and inflation in the early sample period. It is this correlation that the estimates of the Phillips curve pick up. In the last part of the sample, however, the sign of the contemporaneous correlation is negative.

One possible explanation for the falling correlation of inflation and measures of marginal cost is that a smaller share of the variation in inflation now comes from changes in marginal costs (i.e. movements along the Phillips curve) as opposed to other factors (i.e. shifts of the Phillips curve). In this case, estimating κ by concentrating on the economy's response to structural shocks that move the economy along the Phillips curve, for example as pursued by Christiano et al. (2005), may be a preferable approach.[18]

[17] Galí and Gertler (1999) measure inflation on the basis of the GDP deflator while we use the PCE deflator.

[18] Christiano et al. specify an entire general equilibrium model. An alternative approach is to estimate the Phillips curve in isolation from the rest of the model using the information contained in impulse response functions following identified shocks. See Barnichon and Mesters (2020) and Galí and Gambetti (2020).

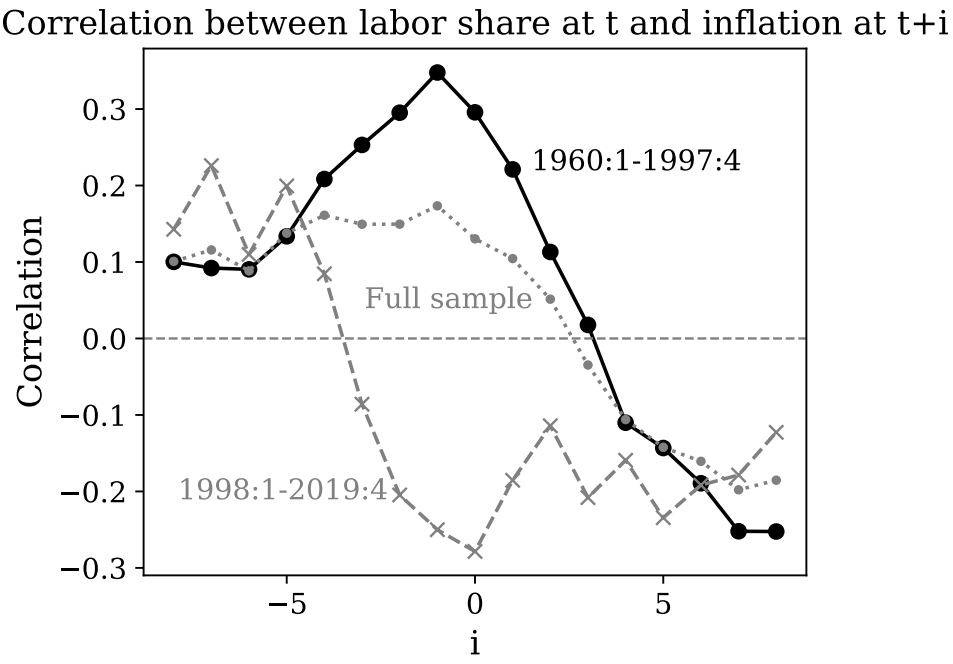


Figure 18.3: Dynamic Correlations

Notes: The figure shows the dynamic correlation function between the labor share and lags and leads of the inflation rate. The data have been HP-filtered.

18.6 Sticky wages and other extensions

The analysis in this chapter has focused on a simple version of the New Keynesian model with sticky prices. This model is useful for illustrating qualitative implications of nominal rigidities but is not a compelling quantitative explanation of the business cycle as we have seen in Figure 18.2. We now describe some of the main extensions of the New Keynesian framework that allow for a richer description of aggregate dynamics. We will pay particular attention to frictions in wage setting as these are a key source of nominal rigidity in addition to price stickiness.

18.6.1 Sticky Wages

The New Keynesian model that we have examined so far includes nominal rigidities in goods prices. This focus on the goods market contrasts with the traditional Keynesian literature that, motivated by the high levels of unemployment observed during the Great Depression, focused on sticky nominal wages.[19] The lack of instantaneous adjustment of nominal wages is consistent with the empirical fact that the distribution of wage changes observed in micro data tend to have a spike at zero, see e.g. Kahn (1997). Sticky wages result in periods when labor is under-utilized thereby providing an explanation for labor market "slack" in the sense that there are workers, although willing to work and actively searching for a job, who appear unable to find a job at the going wage. Sticky wages also hold some theoretical appeal as

[19] A classic reference is Keynes (1936).

the basic sticky-price New Keynesian makes counterfactual predictions for the cyclicality of firm profits.[20]

The Taylor model of overlapping contracts has considerable empirical and intuitive appeal for many parts of the labor market as many wages are adjusted at an annual frequency (see Grigsby et al., 2021). However, as we have argued earlier, the Calvo model offers much analytical convenience. As mentioned above, there may be asymmetries in nominal wage flexibility in terms of the wage change distribution being asymmetric with few downwards movements in wages (of continuing employee-employer relationships). Here, to keep the analysis simple, we will maintain the symmetric modeling of wage stickiness. Wage stickiness may also be an insurance mechanism that can arise in the absence of nominal rigidities. Firms may, for example, be prepared to offer workers insurance against variations in wages because they are better able and willing to absorb profit fluctuations than workers are at handling income variations. Such considerations will typically lead to sticky real wages rather than sticky nominal wages and therefore falls somewhat outside the aims of this chapter.

In the sticky-price model we introduced monopolistic competition in the goods market in order to reconcile the assumption of output being demand determined (in the short run) with firms' willingness to supply goods at a possible predetermined price. A similar assumption is required when modeling sticky wages. The standard avenue taken in the literature is that households rent their labor supply to a continuum of labor unions that differentiate labor and rent it to firms setting a nominal wage that is above the cost that they have acquired it for. Households may therefore sometimes be unable to work the number of hours they would have chosen to in the absence of labor unions and nominal rigidities. However, households are assumed to be the ultimate owners of labor unions and are therefore compensated for this by profits received from these institutions.

In this setup, if, for the sake of discussion, we entertain the idea that wages are sticky and adjusted as in the Calvo model while prices are flexible, the model would be almost identical. In the sticky-price model, the marginal cost of production is the wage (adjusted for productivity) and the wage equals the marginal rate of substitution between consumption and leisure because the household is on its labor supply curve (eq. 18.6). In this setup there is a markup between the marginal cost and the price and due to nominal rigidities this markup can vary in response to shocks. We will now sketch out a sticky-wage model in which prices are equal to marginal costs (production is competitive) but there are frictions in the labor market that introduce a wedge between wages and the marginal rate of substitution between consumption and leisure. This approach to modeling sticky-wages has its roots in Blanchard (1987) and was developed in more detail by Erceg, Henderson, and Levin (2000).

The union sets its wage subject to Calvo-style nominal rigidities and stands ready to supply any amount of labor that is demanded at that wage. A competitive final goods producer combines these varieties of labor to produce a final good selling its output at marginal cost. Let W_t be the nominal wage-index that reflects the cost of a bundle of labor that allows the final goods producer to produce one more unit of goods. Since goods are sold at marginal cost, we have $P_t = W_t$ in all periods. It then follows that price inflation

[20]The sticky-price model predicts that firm profits decline following a positive demand shock. The resulting negative wealth effect on labor supply is actually central to the increase in labor supply with balanced-growth preferences, see Broer, Hansen, Krusell, and Öberg (2020).

will be identical to wage inflation. This model yields the exact same Phillips curve as the sticky-price model. Where the model will differ is its implications for real wages, which are now constant as $w_t = W_t/P_t = 1$. In Section 18.4.1, we argued that the distance between the equilibrium level of output and the efficient level depends on the ratio w_t/A_t. Such a condition no longer holds. Instead, the relevant issue is how the marginal rate of substitution compares to productivity. In the sticky-wage model, households need not be on their labor supply curves and there can be times when they would like to work more, but wages are high and labor demand is therefore "too low," in the sense of being below its efficient (flexible wage) level.

A model with ***both*** sticky wages and sticky prices is, however, somewhat different from the models with only one nominal rigidity. We describe this model in detail in Appendix 18.A.3. As we show there, inflation is now explained by three equations that replace the Phillips curve

$$\pi_t = \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \xi^p \left(\hat{w}_t - \hat{A}_t\right) \tag{18.23}$$

$$\pi_t^w = \beta \mathbb{E}_t \left[\pi_{t+1}^w\right] - \xi^w \left(\hat{w}_t - \hat{A}_t\right) + \kappa^w x_t \tag{18.24}$$

$$\hat{w}_t = \hat{w}_{t-1} + \pi_t^w - \pi_t. \tag{18.25}$$

Eq. (18.23) is the price Phillips curve. It is similar to the standard New Keynesian Phillips curve but now depends on the real wage w_t rather than the output gap. Goods price setters choose their prices taking account of current and future marginal costs, which in this model is just the real wage relative to productivity.[21] Eq. (18.24) is the wage Phillips curve where π_t^w is wage inflation (i.e. the growth rate of nominal wages). Wage setters will increase nominal wages if the marginal rate of substitution is high relative to real wages. Therefore eq. (18.24) is increasing in the output gap (the marginal rate of substitution rises as households work and consume more) and is decreasing in the real wage.[22] This equation is forward looking for the same reason that the price Phillips curve is. Wage setters know their wage could remain fixed for a number of periods so the look ahead to future market conditions. Finally, eq. (18.25) is the log of the identity $w_t = W_t/P_t = (W_{t-1}/P_{t-1})(W_t/W_{t-1})(P_{t-1}/P_t)$. This equation relates the change in the real wage between any two periods to the difference between nominal wage inflation and nominal price inflation.

Figure 18.2 shows results of simulating the model with both sticky wages and sticky prices. There are two things worth noting here. First, we parameterize this model with double the frequency of price and wage adjustments as in the basic model. In the sticky-wage model, prices and wages update once per year on average while in our calibration of the sticky-price model they updated every two years on average. When nominal rigidities layer on top of each other, the pass through from resource utilization to inflation becomes more gradual. Even when they are able to update their prices, intermediate goods firms only raise their prices to the extent their marginal costs rise and the change in their marginal costs is muted by the wage rigidities. Second, note that inflation is more persistent with the

[21] If the model included decreasing returns in production or factors of production other than labor, there could also be an output gap term in eq. (18.23) in addition to the wage term.

[22] Eq. (18.24) is also increasing in productivity because there is a wealth effect on labor supply that raises the marginal rate of substitution between leisure and consumption.

two rigidities. Wage inflation initially falls by more than price inflation leading real wages to fall. Thereafter, the low real wages exert a downward force on price inflation and the real wage only returns to its steady state value gradually.

18.6.2 Other extensions of the basic New Keynesian model

We showed above that the basic New Keynesian model is qualitatively consistent with empirical estimates of the impact of monetary policy shocks. However, as also made clear, quantitatively, the model does not manage to match the data. The same is the case for other structural shocks often studied in macroeconomics such as total factor productivity shocks, shocks to investment efficiency, fiscal shocks, uncertainty shocks, etc. Clearly this basic model fails to capture some important features of business cycle fluctuations. For that reason, much work in the area has considered so-called "medium-scale" models that extend the above framework with the hope of improving the model's quantitative performance. Here we will discuss a few of these extensions and the underlying reason for their introduction into this line of work.

Consumption dynamics: A main feature of many empirical estimates of the macroeconomic impact of aggregate shocks is gradual adjustment over time. We see this above in Figure 18.2 in the hump-shaped response of output to the identified monetary policy shock, but such dynamics are standard findings in the literature also in response to other shocks. There are many ways in which macroeconomists have attempted to model such dynamics.

Under the permanent-income hypothesis, consumption is determined by permanent income and the path of interest rates. It is impossible then to explain a gradual change in consumption without a very particular path for interest rates. To generate a gradual consumption response, some authors replace the preferences in equation (18.1) with a specification such as

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{(C_t - \gamma C_{t-1})^{1-\sigma}}{1-\sigma} - \frac{L_t^{1+\psi}}{1+\psi} \right], \tag{18.26}$$

where $\gamma \in [0, 1)$ and C_{t-1} refers to last period's consumption. These preferences are interpreted as reflecting consumption habits. To see why, note that the term involving consumption can be rewritten as

$$C_t - \gamma C_{t-1} = (1-\gamma)C_t + \gamma(C_t - C_{t-1})$$

i.e. as a weighted average of the level and the change in consumption. Thus, in habit formation models, households are concerned about smoothing both the level and the growth rate of consumption. When γ is large, households effectively worry about smooth growth rates of consumption and in this case the level of consumption will tend to adjust partially to shocks over time. [23]

An alternative way of generating richer dynamics for aggregate consumption is to abandon the complete markets assumptions underlying the permanent income hypothesis. When

[23] This specification is usually referred to as an an internal habit model because the household understands that increasing C_t will affect its utility next period. Another approach treats the habit term as referring to past *aggregate* consumption (taken as given). That approach is called an "external habit" or "catching up with Joneses."

households face uninsurable idiosyncratic risks and borrowing constraints, their consumption choices will tend to reflect current labor market conditions. For example, borrowing-constrained agents will spend strongly out of current income so aggregate consumption will respond more strongly to aggregate disposable income. Similarly, an increase in the risk households face, say an increase in the risk of unemployment, will lead them to cut back on consumption for precautionary reasons. These issues are explored in the literature on heterogeneous-agent New Keynesian models.

Capital, investment, and adjustment costs: The model discussed so far has only one factor input: labor. Quantitative models typically also include capital accumulation both because capital is important as a savings vehicle and also because investment demand is an important part of aggregate demand. Typically, one assumes a Cobb-Douglas technology:

$$y_{j,t} = A_t k_{j,t}^{\alpha} \ell_{j,t}^{1-\alpha}. \tag{18.27}$$

Let's assume that capital is owned by households and rented out to firms at the (real) rental rate r_t^k. Now assume that capital accumulates over time according to a standard neoclassical specification:

$$K_{t+1} = (1-\delta)K_t + I_t,$$

Unfortunately, this model implies that investment demand becomes extremely sensitive to variations in monetary policy. To see this, note that firms will continuously adjust their capital demand so as to equate the marginal product of capital with the cost of capital $r_t^k + \delta$. In steady state, investment satisfies $I_t = \delta K_t$ and with a small δ, the flow of investment is small relative to the capital stock. Small percentage changes in the desired capital stock then translate to large percentage changes in investment. The model as written would then imply that small changes in interest rates lead to large changes in investment and equilibrium output.

To limit such high investment volatility, it is common in medium-scale models to include sources of adjustment costs. A common specification is:

$$K_{t+1} = (1-\delta)K_t + \xi\left(\frac{I_t}{K_t}\right)K_t$$

where the function $\xi(I_t/K_t)$ is assumed to be increasing but concave and captures adjustment costs.[24] Notice here that, because of concavity, it is costly to vary the investment rate. This specification is therefore able to generate gradual adjustments in the flow of investment and hump-shaped aggregate dynamics.

Empirically, the slope of the Phillips curve is low relative to what we would expect based on the rigidity of prices as measured in micro-data as we discussed in Section 18.5.2. This tension is exacerbated in the New Keynesian model with capital accumulation. In the short-run, the capital stock is predetermined and the only way to produce more in the aggregate is to use more labor. If we use the production function (18.27), the elasticity of output with respect to labor is $1-\alpha$, which means we need to use $1/(1-\alpha) > 1$ units of labor in order to produce one more unit of output. Thus, relative to the specification in (18.3), marginal costs

[24] Typically one assumes $\xi > 0, \xi' > 0, \xi'' < 0$, and that $\xi(\delta) = \delta$ so that Tobin's Q equals 1 in the deterministic steady state.

are more sensitive to the quantity produced and the Phillips curve becomes steeper. One way to address this concern is to allow for variable capacity utilization whereby the effective capital services are not predetermined giving the economy an additional opportunity to adjust production beyond changes in labor effort.

Chapter 19

Credit market frictions

Vincenzo Quadrini

19.1 Introduction

The study of the role of financial market frictions for the dynamics of the macro-economy has a long tradition that started well before the 2008 financial crisis. Bernanke and Gertler (1989) and Kiyotaki and Moore (1997) are two well-known studies that incorporate financial frictions in general equilibrium models and became the standard references for this literature. However, before the 2008 financial crisis, the mainstream approach to the study of the macro-economy abstracted from financial frictions. This was, in part, motivated by the view that, although markets are clearly incomplete, the importance of financial frictions for the dynamics of the macro-economy is somewhat negligible. Based on this view, it became preferable to use models with complete markets because they allow for simpler analytical characterization. These models featured many frictions such as sticky prices, sticky wages, adjustment costs in investment, variable capital utilization, matching frictions in the labor market, but not financial frictions.

The view that financial frictions were somewhat negligible for understanding the macro-economy is surprising considering that one of the largest recessions in modern times—the 1929 Great Depression—was associated with extensive financial issues such as a deep banking crisis. As discussed in the first chapter of this book, the Great Depression was very important for the subsequent research in macroeconomics. However, some prominent scholars did emphasize the centrality of financial markets for understanding the macroeconomic dynamics of the Great Depression (see, for example, Calomiris, 1993 and Bernanke, 2024), but the role of financial frictions remained limited in many macroeconomic studies. An example is Kehoe, Prescott, et al. (2002), a book collecting contributions from distinguished macroeconomists on the topic of the Great Depression. None of the book chapters explored the role played by financial frictions for understanding the macroeconomic implications of the Great Depression.

The 2008 financial crisis also known as the Great Recession, represents a turning point. The crisis made clear that macroeconomic models constructed on the assumption of friction-less financial markets were missing important elements for understanding the dynamics of the economy. These models needed to be extended to have a more prominent role played by the financial sector and the goal of this chapter is to illustrate one way to do that.

Given the introductory nature of the chapter, we start with the most standard model

used in macroeconomics, the neoclassical growth model. The neoclassical growth model has been characterized in previous chapters but under the assumption of complete markets. Here we relax this assumption and introduce a special form of financial frictions. Although the extension with incomplete markets introduces only a special form of financial frictions, some of the basic properties illustrated in this chapter are quite general and shared by other models used in the literature. Before describing the specific modeling, however, it would be helpful to provide an empirical motivation for why the joint analysis of financial and real markets deserves attention for understanding the dynamics of the macro-economy.

19.2 Financial and real markets

Why should macroeconomists pay attention to the joint dynamics of financial and real markets? Besides the anecdotal observation that financial markets could have played some role in specific episodes of booms and busts around the world, a more systematic observation is that credit flows are highly pro-cyclical. The pro-cyclicality is evident not only at the business cycle frequency but also, and perhaps more importantly, over the medium term.

The top panel of Figure 19.1 shows that changes in U.S. credit market liabilities move closely with the U.S. cycle. In particular, debt growth drops significantly during recessions. There are exceptions. For example, credit in the household sector did not contract in the 2001 recession, contrary to business debt. The drop in credit was especially large during the 2008 crisis. The procyclicality of corporate debt has also been shown in many studies using micro data from publicly traded firms.[1]

We can also see the cyclical properties of financial markets from more direct indicators of credit tightening based on survey data. The bottom section of Figure 19.1 plots, for the United States, the net percentage of senior bank officers reporting tightening of credit standards for commercial and industrial loans, and for credit cards. A higher credit standard index indicates that it is more difficult to get a loan from a bank. The figure shows that U.S. banks tighten their credit standards during recessions.

Other indicators of credit tightening such as credit spreads—the interest rate differential on corporate bonds issued by companies with weak credit rating over the interest paid by government bonds—convey a similar message. As shown in the top panel of Figure 19.2, interest rate spreads tend to increase right before a recession. This suggests a deterioration of the credit capacity of certain businesses, making more difficult for them to borrow.[2]

The bottom panel of Figure 19.2 plots the flow of new debt in the U.S. private sector and the U.S. unemployment rate. The graph illustrates the strong negative co-movement between the growth of debt and unemployment. This is another illustration of the strong linkage between real and financial sectors.

Why is the co-movement in real and financial variables relevant? In a world in which markets are complete, which is the case in the standard neoclassical model, the financial structure of households and firms would not necessarily follow a pro-cyclical pattern. For example, since in a recession there are more unemployed workers, the household sector could borrow more to fund the consumption of unemployed workers. This could lead to a

[1] See Covas and denHaan (2011) and Begenau and Salomao (2019).

[2] See also Gilchrist, Yankov, and Zakrajsek (2009).

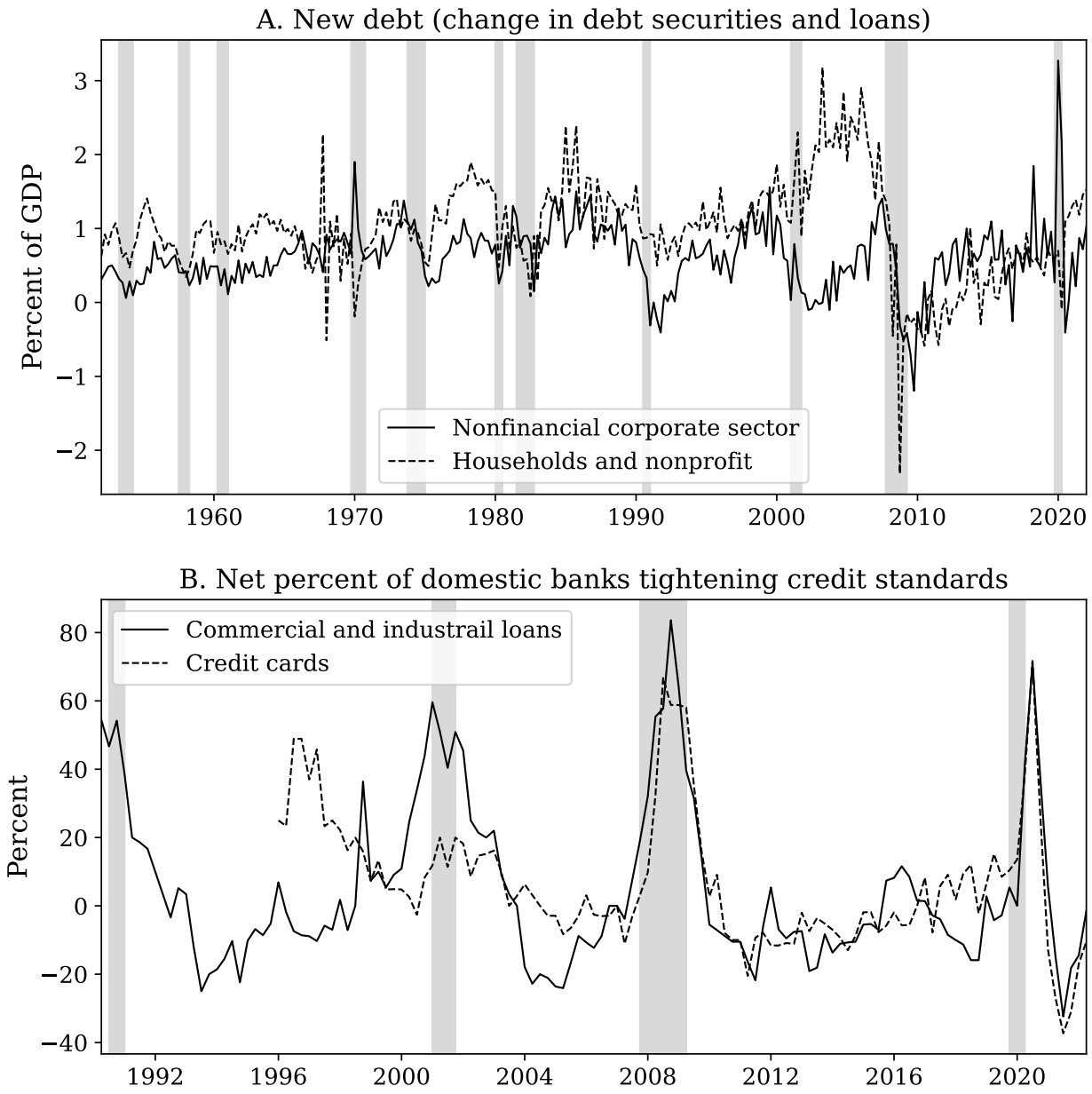


Figure 19.1: **Panel A**: Change in the volume of U.S. credit market instruments in the households and business sector as a percentage of GDP. **Panel B**: Index of credit tightening in commercial and industrial loans and in credit cards.

counter-cyclical flows of new debt. The fact that credit flows are pro-cyclical and the index of tightening standards is counter-cyclical suggests that the complete-market paradigm has limitations in capturing the joint dynamics of real and financial variables. This is especially true for the index of credit tightening: if markets were complete, there is no reason for lenders to change their 'credit standards' over the business cycle.

However, the fact that there is a strong co-movement between financial and real flows—as shown in Figure 19.1 and the bottom panel of Figure 19.2—does not tell us anything about

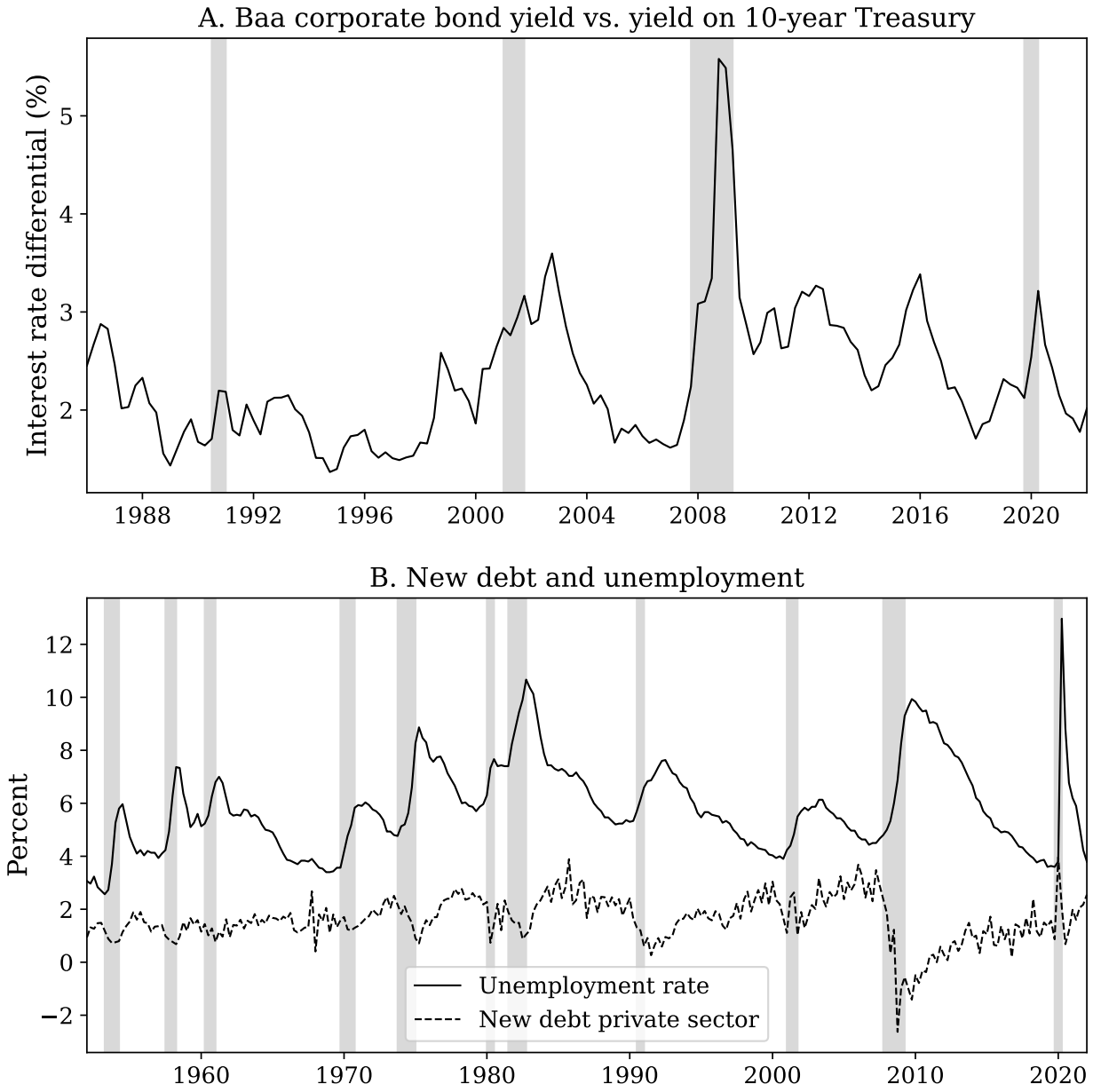


Figure 19.2: **Panel A**: Interest rate differential between U.S. corporate bonds issued by companies with a Baa credit rating and the yield on 10-year Treasury notes. **Panel B**: Change in the volume of credit market instruments in the U.S. private sector (households and businesses) as a percentage of GDP and U.S. unemployment rate as a percentage of the labor force.

possible causal links: does lower credit growth cause recessions or do recessions cause lower credit growth? Conceptually, we could have three possible scenarios:

1. *Real activity causes movements in financial flows.* It is possible that the contraction in consumption and investment expenditures, as well as employment, is the direct response to changes in real factors such as productivity. In this case, borrowers cut

their debt simply because they need fewer funds to conduct economic transactions and to finance their expenditures. If this were the only linkage between real and financial flows, the explicit modeling of the financial sector would be of limited relevance for understanding movements in real economic activity. The neoclassical model is built on this assumption.

2. *Propagation.* A second possibility is that non-financial factors, such as changes in productivity, are the initial driving forces for movements in economic activity. However, financial factors affect how these initial changes propagate to the real sector of the economy. For example, as investment and employment both fall in response to a decline in productivity, the credit capacity of borrowers also deteriorates. This could happen, for instance, if the fall in investment generates a fall in the market value of assets that are used as collateral. The deterioration in accessible credit forces borrowers to reduce investment and hiring more than they would have in absence of the credit contraction. Thus, financial frictions could *amplify* the macroeconomic impact of non-financial shocks.

   Although in this example financial frictions amplify the shock, they could also dampen it. For example, an improvement in technology stimulates investment. The presence of financial frictions, however, could limit the investment growth because the credit required to finance the desired investment boom could be unobtainable. In general, whether financial frictions act as amplification or dampening mechanisms, it is plausible to think that they impact, somewhat, the propagation of non-financial shocks. The question, then, is whether their impact is quantitatively important.

3. *Financial shocks.* A third possibility is that the initial disruption or shock arises directly in the financial sector. After hitting the financial sector, the shock propagates to the real side of the economy. For example, a disruption in financial markets could make it more difficult to transfer funds from lenders to borrowers causing a contraction in consumption and investment spending and leading to a decline in employment. This type of disruption is now commonly referred to as 'credit' or 'financial' shock. In reduced form, a financial shock could be modeled as a change in the parameter that determines the severity of the financial frictions in the model.

Prior to the 2008 financial crisis, a large majority of studies in the macro-finance literature focused on the second possibility, that is, on the role of financial frictions for the 'propagation' of non-financial shocks.[3] The main idea was that financial frictions could make the real impact of a shock bigger (amplification) or smaller (dampening). However, financial frictions are not the initial 'cause' of macroeconomic expansions and contractions: something else has to happen first in the nonfinancial sector. The analysis of financial shocks as a 'source' of macroeconomic fluctuations has received more attention after the 2008 financial crisis.

This chapter will present some parsimonious ways to formalize financial frictions and show how they affect the propagation of non-financial shocks and of financial shocks to the rest of the economy. The chapter will not discuss at length the first possibility because,

[3]This was the case for Bernanke and Gertler (1989) and Kiyotaki and Moore (1997) and the related literature.

as observed above, if this is the most relevant linkage between real and financial flows, the explicit modeling of the financial sector will be of limited relevance for understanding the dynamics of the real economy.

19.3 Modeling financial frictions

Financial frictions arise when certain trades in financial assets or claims are not available. In the context of an Arrow-Debreu model, they emerge when trades in certain contingencies cannot take place. This limits the agents' ability to shift spending across time or states of nature. The absence of trading markets becomes relevant only if agents are heterogeneous in some dimension that induces a reason to trade. Thus, all models with relevant financial frictions share two features:

1. *Missing markets*: Markets for the trade of certain claims are absent.

2. *Heterogeneity*: Agents are heterogeneous in some important dimension.

These two characteristics are necessary for financial frictions to be relevant but are not sufficient. The absence of some markets could be irrelevant if, in equilibrium, agents would choose not to trade in these markets anyway. Thus, market incompleteness and heterogeneity must interact for financial frictions to be relevant for the economy. The literature has considered various forms of market incompleteness and heterogeneity.

19.3.1 Missing markets

To demonstrate the role of missing markets, it is useful to start with a counter-example known as the Modigliani and Miller (1958) principle. The principle states that, with complete markets, the sources of funding (for example debt versus equity) do not matter for real decisions. For example, whether an agent funds investment with debt or equity does not affect how much the agent invests—the agent is indifferent between the various sources of funding. But when we talk about financial frictions, we deal with environments where agents are not indifferent between the various sources of funding. Thus, if we want to consider these types of environments, we need to depart from the frictionless setting.

Modigliani-Miller principle. To illustrate the Modigliani-Miller principle, we use a version of the complete market model studied in Chapter 7. A firm purchases capital k at time 0, at price 1 (one unit of capital at time 0 is worth 1 unit of time-0 consumption goods). Then in period 1, the state i (shock) is realized, after which the firm hires labor ℓ_i at price $p_i w_i$. Here p_i is the Arrow-Debreu price (consumption goods paid in period 0 for the delivery of one consumption good in period 1) and w_i is the price of one unit of labor in terms of consumption goods paid in period 1. Thus, $p_i w_i$ is the wage in terms of consumption goods in period 0. The firm produces $F(z_i, k, \ell_i)$ at date 1 and each unit of output is worth p_i units of consumption goods in period 0. Finally, the firm sells the undepreciated capital $(1-\delta)k$, each unit worth p_i units of period-0 consumption.

The firm solves the problem

$$\max_{k,\{\ell_i\}_i} \Big\{ -k + \sum_i p_i \Big[F(z_i, k, \ell_i) - \ell_i w_i + (1-\delta)k \Big] \Big\}. \tag{19.1}$$

This problem does not specify how the firm finances the purchase of capital k in period 0. So let's assume that the firm has two choices: debt, denoted by b, or equity, denoted by e. In the first case the firm will borrow while in the second the firm sells shares to future dividend payments. Given these two sources of funding, we have that $k = b + e$.

Denote by R the risk-free gross interest rate paid on the debt. Then, the value of dividends in terms of period-0 consumption paid by the firm in period 1 at state i are $d_i = p_i[F(z_i, k, \ell_i) - \ell_i w_i + (1-\delta)k - Rb]$. Since the firm maximizes the value for the providers of equity, the shareholders, the optimization problem is now

$$\max_{k,\{\ell_i\}_i} \Big\{ -(k-b) + \sum_i p_i \Big[F(z_i, k, \ell_i) - \ell_i w_i + (1-\delta)k - Rb \Big] \Big\}. \tag{19.2}$$

The firm's owners pay $e = k - b$ at time 0, and then in period 1 they receive the profits after the repayment of the debt inclusive of the interests, Rb.

As we have seen in Chapter 7, in a complete markets equilibrium, prices satisfy $\sum_i p_i R = 1$. This implies that the variable b cancels out in Problem 19.2 and we get back to the previous Problem 19.1 which does not specify the sources of funding. Thus, the choice of the financial structure—b versus e—is indeterminate.

The argument is very general and applies to any type of financial instruments. For example, we can consider short-term vs. long-term debt. The only requirement is that the various instruments are priced competitively, which is the case in an Arrow-Debreu economy. The argument could also apply with incomplete markets, if the various financial instruments are priced in the same way firms discount future dividends. In this chapter, however, we will consider environments in which this is not the case.

Modelling missing markets. The approaches used in the literature to formalize missing markets and possibly leading to the violation of the irrelevance principle of Modigliani and Miller, can be divided in two categories: 'exogenous' and 'endogenous'.

1. *Exogenous market incompleteness.* In this category we include models that impose, by assumption, that certain claims or assets cannot be traded. A common assumption is that agents can save or borrow with non-contingent bonds. However, they cannot purchase or issue bonds with payoffs that are contingent on information that will be revealed in the future. In general, the goal of these studies is not to explain why certain assets cannot be traded but to understand the consequence of the missing markets.

   These studies start from the observation that in the real world a large volume of financing transactions is in the form of borrowing and lending with fixed repayment (standard debt contracts). Of course, there are also financial contracts with payouts contingent on future events. However, the volume of these contracts is smaller than the theory would predict. The modeling assumption that borrowing can be done

only with non-contingent debt is a pragmatic approximation to the complexity of the actual economy. Another common assumption is that there is a limit to the debt that an individual agent can take (an exogenous borrowing constraint). The limit is not necessarily fixed and could depend on the characteristics of the borrower. However, there is no attempt to motivate with explicit micro-foundations why borrowing is limited.

2. *Endogenous market incompleteness.* Models with endogenous market incompleteness assume that the set of feasible contracts is restricted by incentive considerations. Markets are missing because parties are not willing to engage in certain trades as a result of agency frictions, which typically take one of two forms:

   (a) *Information asymmetry.* In many instances, borrowers have more information than lenders. For example, if repayment is contingent on the performance of the business and that performance depends on 'unobservable' effort from the borrower, the borrower has an incentive to choose lower effort. Since effort is observed only by the borrower, shirking cannot be detected. This creates a moral hazard problem. In some contexts, information asymmetry gives rise to adverse selection: since the riskiness of borrowers cannot be detected by lenders, those applying for loans tend to be riskier borrowers. Adverse selection is thought to be an important issue also in insurance markets.

   (b) *Limited enforcement.* Lenders may have the same information as borrowers and they can observe whether borrowers deviate from the agreed action (for example the amount that should be repaid upon the realization of a specific event). However, lenders are unable to force borrowers to fulfill their obligations. If the amount repaid is lower than initially agreed, there is nothing that lenders can do to enforce the agreed repayment.

   Although models with limited enforcement are typically easier to characterize analytically than models with information asymmetries, they share a common property: the higher is the net worth of borrowers, the higher are the (incentive-compatible) funds that can be raised externally. This is also a property of many models with 'exogenous' market incompleteness. Therefore, for a variety of questions often addressed in macroeconomics, whether market incompleteness is exogenous or endogenous does not make a big difference. However, there are questions for which it does matter. Also, an advantage of models where market incompleteness is endogenous, is that they are less vulnerable to the Lucas critique, that is, the fact that certain parameters may not be invariant to policy changes. A simple example is a policy that increases the penalty for defaulting on the debt. Because of the higher penalty, creditors anticipate that borrowers have less incentive to renege on the debt obligations and, as a result, they are willing to lend more. This, effectively, relaxes the borrowing constraint, something that would not be captured in a model where the borrowing constraint is exogenous. Bernanke and Gertler (1989) and Kiyotaki and Moore (1997) are examples of endogenous borrowing constraints: the first based on information asymmetry and the second on limited enforcement.

19.3.2 Heterogeneity

There are different approaches used in the literature to incorporate heterogeneity. In some models, agents are ex-ante identical but they become heterogeneous ex-post as a result of uninsurable idiosyncratic shocks, as in the Bewley-Huggett-Aiyagari economies already studied in Chapter 11. Because incomplete markets generate a complex degree of heterogeneity in this model, many applications using Bewley-Huggett-Aiyagari economies abstract from aggregate uncertainty with few exceptions.[4] Instead, the majority of studies that investigate the importance of financial frictions in the presence of aggregate shocks have used alternative modeling approaches where heterogeneity is less complex.

A popular approach is the assumption that there are only two types of agents that are ex-ante and permanently different in preferences and/or technology. In equilibrium, one type of agents borrows while the other lends. In some cases, there is heterogeneity also within the same type of agents. However, the aggregate behavior of each type can be characterized as if there is a representative agent. This is because the model allows for linear aggregation of all agents belonging to the same type (but not between different types).[5].

In many models, financial frictions are important because they limit the reallocation of resources from agents that are less productive to agents that are more productive. There is an incentive for those who will be more productive, to save in order to overcome future borrowing constraints. Over time, these agents may accumulate enough wealth so that they are no longer dependent on external funding. To maintain the significance of financial frictions in the long run, further assumptions are needed. The box provides a non-exhaustive list of assumptions made by various studies in the literature.

Assumptions that prevent self financing

Finite life span. Some models assume that borrowers are finitely lived. Examples include models with overlapping generations where newborn agents have no initial wealth. Over time, they accumulate wealth and become unconstrained. However, since there are agents that are newly born in every period, at any point in time there are always some agents that face binding financial constraints. A similar mechanism arises in models with an industry dynamic structure where exiting firms are replaced by new entrant firms.

Different discounting. Another popular approach is to assume that agents are infinitely lived but some of them discount the future more than other agents. More impatient agents end up being the borrowers while patient agents lend. For impatient agents, the cost of external financing—the interest rate—is lower than their inter-temporal discount rate. Therefore, they do not save enough to reach a point at which the borrowing constraint is no longer binding in long-run. This implies that in every period some agents are always constrained.

Tax benefits. A similar outcome can be obtained with tax benefits of debt. For example,

[4] One of the first exception is Krusell and Smith (1998). Others include Cooley, Marimon, and Quadrini (2004), Guerrieri and Lorenzoni (2010), Khan and Thomas (2011) and Arellano, Bai, and Kehoe (2019).

[5] Carlstrom and Fuerst (1995) and Bernanke, Gertler, and Gilchrist (1999) are examples of this approach.

the tax deductibility of interest payments from corporate earnings generates a preference for debt over equity (higher leverage). In this case the two types of agents are, typically, households (who do not get a tax benefit from debt) and firms (who benefit from the debt shield). This assumption is especially popular in structural finance with various applications to macroeconomics.

Convenience yield. An assumption that is becoming more popular in the macro-finance literature is that debt issued by certain agents provides liquidity services to other agents. Sometimes this is referred to as a 'convenience yield.' This may be because this type of debt is safe and can be easily liquidated and accepted as collateral in emergencies, unlike other types of debt. Another way to interpret this idea is that this particular type of debt provides utility flows. Because of these utility flows, agents are willing to hold it even if the interest rate is lower than the inter-temporal discount rate. Although the most common application of the convenience yield is for bonds issued by governments, it can be extended to other types of debt.

Wage bargaining. A further assumption considered in the literature is that external financing (debt or outside equity) is preferred to inside financing (entrepreneurial equity) because it affects the bargaining position of firms in the negotiation of wages and/or executive compensation. If the compensation of workers and/or managers is determined through some form of bargaining (perhaps through labor unions in the case of workers), highly levered firms would be able to negotiate lower wages because the debt reduces the bargaining surplus. This mechanism introduces an incentive for firms to take on more debt and can leave them financially constrained.

This section has outlined the most popular approaches used in the literature to formalize financial frictions. Of course, it is not possible to illustrate all of them in detail in a single chapter. We will then focus on a particular formalization. However, the most salient properties that will be outlined in the next sections are similar to those characterized in models that adopt a different formalization of financial frictions.

19.4 Adding financial frictions to the neoclassical model

The starting point is the neoclassical growth model studied in previous chapters. In that model, markets are complete and there is only one representative agent. Here we maintain the assumption that there is a representative household that has the same characteristics as in the standard model. However, we place some more structure on the operation of firms. More specifically, we assume that capital accumulation (investment) is chosen by firms, not households. This implies that firms solve a dynamic problem because the investment made today increases profits in the future.

Households continue to be the owners of the firms, which operate to serve the interests of households. However, we will think of the firms as distinct from the households and they represent a second type of agent. Thus, we introduce heterogeneity by distinguishing households from firms. Based on assumptions that we will introduce gradually in the model, we will see that, in equilibrium, firms borrow from households and households lend to firms.

The objective of firms is to maximize the current value of dividends. The assumption that households are the owners of firms, which operate in the interest of households, implies the discount factor used by firms is the same discount factor used by households to discount future payments. As they use the same discount factor, firms make the same production and investment decisions that would be made directly by households.

For a household, a unit increase in consumption at time $t > 0$ has a utility value of $u_1(C_t, L_t)$, where $u_i(\cdot, \cdot)$ represents the derivative with respect to the ith argument. Thus, $u_1(C_t, L_t)$ is the marginal utility of consumption at time t. The discounted value at time zero of that utility is $\beta^t u_1(C_t, L_t)$. We now ask the following question: What should be the increase in consumption at time zero that gives the same flow of utility as one unit increase in consumption at time t?

Denote by Δ the marginal increase in consumption at time zero. The extra utility received at time zero from this increase is $\Delta \cdot u_1(C_0, L_0)$, that is, the increase in consumption multiplied by the marginal utility of consumption at time zero. The increase in consumption Δ that provides the same utility as the unitary increase in consumption at time t is determined by solving the condition $\Delta \cdot u_1(C_0, L_0) = \beta^t u_1(C_t, L_t)$. Rearranging we obtain

$$\Delta = \frac{\beta^t u_1(C_t, L_t)}{u_1(C_0, L_0)}.$$

Thus, a unit of consumption at time t is equivalent to $\beta^t u_1(C_t, L_t)/u_1(C_0, L_0)$ units of consumption at time zero. The term $\beta^t u_1(C_t, L_t)/u_1(C_0, L_0)$ is known in the literature as the 'stochastic discount factor.' Because the stochastic discount factor captures how households—the owners of firms—evaluate future payments in the current period, firms use this factor to discount dividends that will be paid in the future.

For notational convenience we denote the stochastic discount factor using the variable $m_t = \beta^t u_1(C_t, L_t)/u_1(C_0, L_0)$, a notation often used in the field of finance. The firm's objective function, then, can be expressed as

$$\mathbb{E} \sum_{t=0}^{\infty} m_t d_t,$$

where d_t are the dividends paid by the firm at time t. The objective of the firm is to maximize the expected discounted value of dividends using m_t to discount dividends paid in the future.

The next step is to place some structure on the financing decisions of firms. Since capital is accumulated by firms, they need to finance investment. We allow for two forms of financing: equity and non-contingent debt. Equity refers to the households' claims to the dividends of the firm. Non-contingent debt is the amount borrowed from households. If the firm borrows b_{t+1}/R_t at time t, it promises to repay b_{t+1} units at time $t + 1$. The variable R_t is the gross interest rate (one plus the interest rate), while its inverse, $1/R_t$, is the price of the debt. The interest rate is determined in general equilibrium to clear the credit market.

So far, even though we added more structure to the neoclassical growth model by assuming that capital is accumulated by firms and funded with debt and equity, this is inconsequential for the equilibrium allocation. Proposition 19.1 introduced below establishes that, in the absence of additional assumptions, the equilibrium debt chosen by firms is indeterminate and the real allocation is the same as in the standard neoclassical growth model.

This is an example of the Modigliani-Miller's irrelevance principle. To make the financial decision relevant, we need to introduce some frictions which we do with a set of assumptions. Assumptions (19.4.1)-(19.4.3) differentiate the model with financial frictions presented here from the standard neoclassical growth model.

The first assumption is that debt financing is cheaper than equity financing.

Assumption 19.4.1 *The cost of debt for firms is lower than the cost of equity.*

One way in which we can make this assumption operational is with a tax benefit of debt. Even if the gross interest rate paid by the firm is R_t, it receives an indirect tax benefit due to the tax deductability of interest payments from corporate profits. The tax benefit is equivalent to a subsidy $\tau > 0$ so that the effective gross interest rate paid by the firm is

$$\tilde{R}_t = R_t/(1+\tau).$$

This assumption has been used by several studies in structural corporate finance (see, for example, Hennessy and Whited, 2007) and macro-finance (see, for example, Jermann and Quadrini, 2012). However, this is not the only way to generate a cheaper cost of borrowing for the firm. Another approach is the assumption that the debt has a convenience yield for households. This can be implemented by assuming that the firm's debt enters the utility of households. Thanks to the flow of utility, households are willing to hold the debt even if the interest rate is lower than the inter-temporal discount rate $1/\beta - 1$.

As we will show below, Assumption 19.4.1 implies that firms prefer debt over equity: they raise debt and pay out dividends. Thus, the tax benefit breaks the indifference property of the financial decisions of firms and introduces a pecking order in the source of funds. This also implies that firms would like to issue an infinite amount of debt. To prevent excessive borrowing, then, it is customary to impose a borrowing limit. We follow the same approach here.

Assumption 19.4.2 *Given the price of capital, p_t, the debt issued by the firm is subject to the borrowing constraint,*

$$\frac{b_{t+1}}{\tilde{R}_t} \leq \xi p_t k_{t+1}.$$

According to the constraint, the debt cannot be larger than a fraction ξ of the value of capital. Capital acts as collateral and effectively limits the amount of debt that firms can issue since only a fraction ξ of the value of capital can be funded with debt. The remaining fraction, $1-\xi$, must be funded with equity, that is, funds that are owned directly by the firm and, indirectly, by households. Collateralized credit is a very common practice in bank lending. Although we do not attempt to justify the borrowing constraint with some deeper theory, it can be derived from limited enforcement assumptions.[6]

[6]For example, if in the eventuality of default the lender will only be able to confiscate a fraction ξ of capital, the recovered value will be $\xi p_t k_{t+1}$. Knowing this, the borrower could default whenever the debt exceeds this value. But then the lender will not be willing to lend more than this value.

The price of capital p_t plays an important role since its fluctuations affect the firm's access to credit. In our model, however, p_t is always one since it is always possible to convert one unit of capital into a unit of consumption. In many studies, however, it is assumed that the conversion of capital to consumption goods or vice versa, is not one-to-one. This is the case if there are adjustment costs: if we want to increase the stock of capital by one unit, we need to give up more than one unit of consumption. Furthermore, as we add more and more capital, the needed units of consumption increase. This is the idea of convex capital adjustment costs. In this case the price of capital p_t is given by the marginal cost, which increases with investment. For example, convex adjustment costs are present in the financial accelerator model of Bernanke et al. (1999). When the adjustment cost is prohibitively high, the stock of capital is constant and we have, effectively, an economy with a non-reproducible asset similar to land. This is the assumption made in Kiyotaki and Moore (1997). In both papers, the endogenous change in the price of capital plays an important role because it affects the agent's ability to borrow.

To keep the analysis of this chapter simple, we do not attempt to introduce capital adjustment costs to make the price of capital move endogenously. Instead, we will make a very extreme simplification: we assume that the price of capital p_t is exogenous and depends only on aggregate shocks. Of course, this is a very extreme simplification since prices are endogenous objects and they are determined in equilibrium. The simplification, however, allows us to derive intuitive analytical expressions.

The lower cost of debt, paired with a binding borrowing constraint (Assumptions 19.4.1 and 19.4.2), already introduce a wedge that distorts real allocations. However, this particular wedge plays a relatively limited role in affecting the response of the economy to aggregate shocks. To understand why, suppose that the economy experiences an increase in productivity or TFP. Also, suppose that the higher TFP is expected to persist for several periods in the future, so that firms have an incentive to increase investment. Assumption 19.4.1 implies that the most attractive source of funding for investment is debt since it is cheaper than equity. However, because of the borrowing constraint, only part of the increase in investment can be funded with debt. The other part needs to be funded with equity. In practice, equity financing involves paying lower dividends. Dividends could be negative, which in the model is equivalent to new equity issuance (new shares sold to households). Since there is no cost to adjusting dividends, the firm can always obtain the extra funds by issuing equity. Thus, the borrowing constraint imposes a limit only on one source of funds (debt) and it does not restrict the other source of funding (equity). By having unlimited access to equity, even if more costly, the firm is able to fund the desired level of investment. The firm retains sufficient financial flexibility that its response to shocks is only marginally affected by the frictions set by the borrowing limit. For the financial frictions to have a more substantial role, we also need to limit the ability to use other sources of funding. In the model we need to restrict access to equity financing.

One way to limit the use of equity financing is to impose that the firm cannot pay negative dividends. The firm can re-invest the current profits by paying zero dividends but it cannot issue new equity (sell new shares). This approach has been used in many models proposed in the macro and corporate finance literatures. Here we take a more general approach and make the following assumption.

Assumption 19.4.3 *Firms have a target ϕ for the payment of dividends and incur the following cost if they deviate from the target*

$$\kappa(d_t - \phi)^2.$$

The convexity assumption could be justified by several considerations. When dividends become negative, it means that the firm raises external equity by issuing new shares. Hansen and Torregrosa (1992) and Altinkilic and Hansen (2000) have shown that underwriting fees display increasing marginal cost in the size of the offering. We can also think of the adjustment cost as capturing the preferences of managers for dividend smoothing as shown by Lintner (1956) and confirmed by subsequent studies. This could be the reflection of agency conflicts which induce managers to keep dividend payment stable, something that we do not model explicitly. Instead, we recognize that in reality firms implement dividend policies that are smoother than their corporate earnings and we capture that in reduced form by assuming a quadratic cost when deviating from the dividend target ϕ.

19.4.1 Optimality conditions for firms

We now have all the elements to characterize the problems solved by firms and households. The firm's optimization problem can be written recursively as

$$\Omega(S;k,b) = \max_{d,\ell,k',b'} \left\{ d + \mathbb{E}m'\Omega(S';k',b') \right\} \tag{19.3}$$

subject to

$$F(z,k,\ell) - w\ell + \frac{b'}{\tilde{R}} = b + p\Big[k' - (1-\delta)k\Big] + \varphi(d)$$

and

$$\xi p k' \geq \frac{b'}{\tilde{R}},$$

where we have used the function $\varphi(d) = d + \kappa(d-\phi)^2$ to denote the dividend payment to households (shareholders) plus the cost of deviating from the targeted dividend. To pay d to shareholders, the firm uses $\varphi(d)$ units of resources. The function $\Omega(S;k,b)$ is the expected discounted value of dividends. The firm's value depends on the aggregate state, S, and on the individual states given by the beginning-of-period stock of capital, k, and the debt inherited from the previous period, b. The value of the firm is given by the current dividend, d, plus the expected next period value of the firm, $\mathbb{E}m'\Omega(S';k',b')$. The discount factor m' is not affected by the firm's policy. This is because the firm is atomistic and its impact on aggregate variables is negligible. The aggregation of the policies chosen by all firms, however, do affect m'.

The problem is subject to two constraints. The first is the budget constraint. On the left-hand side we have the sources of funds: the firm's operational profits, $F(z,k,\ell) - w\ell$, and the funds raised with the new issuance of debt, $b'/\tilde{R}$. On the right-hand side we have the uses of funds: the repayment of the outstanding debt, b, capital investment, $p[k' - (1-\delta)k]$, and the resources needed to pay dividends, $\varphi(d)$.

The second constraint to the firm's problem is the borrowing limit. This imposes an upper bound to the capital that can be funded with debt: The left-hand side is the collateral value of capital, while the right-hand side represents the funds raised with new debt.

To characterize the optimal policies of the firm, we derive the first order conditions with respect to dividends, d; labor, ℓ; new stock of capital, k'; and new debt, b'. The conditions lead to[7]

$$F_3(z, k, \ell) = w, \tag{19.4}$$

$$\mathbb{E}m' \left(\frac{\varphi'(d)}{\varphi'(d')} \right) \Big[(1 - \delta)p' + F_2(z', k', \ell') \Big] = (1 - \mu\xi)p, \tag{19.5}$$

and

$$\mathbb{E}m' \left(\frac{\varphi'(d)}{\varphi'(d')} \right) \tilde{R} = 1 - \mu, \tag{19.6}$$

where μ is the Lagrange multiplier associated with the borrowing constraint, $F_i(\cdot, \cdot, \cdot)$ represents the derivative with respect to the ith argument, and $\varphi'(\cdot)$ represents the first derivative.

The optimality condition for labor, equation (19.4), equalizes the marginal product of labor to the wage rate. This is the same condition as in the standard neoclassical growth model. Therefore, financial frictions do not distort directly the optimal hiring of labor. This is not the case, however, for capital investment.

The optimality condition for investment, equation (19.5), equalizes the effective cost of buying an extra unit of capital (the right-hand side term) to its expected discounted gross return (the term on the left-hand side). If the borrowing constraint is binding, implying that the multiplier μ is positive, the effective cost of an additional unit of capital is lower than its price. The effective cost is lower because the extra unit of capital allows the firm to borrow more since capital can be used as a collateral. Provided that the cost of debt is lower than the cost of equity—Assumption 19.4.1—this is a benefit for the firm and reduces the effective cost of capital. Notice that the benefit increases with the parameter ξ because a higher ξ allows a unit of capital to be funded with more debt, which is cheaper. In the literature, the collateral benefit of capital is often referred to as 'collateral premium.' Agents are willing to hold collateralizable capital even if its direct return is lower than other assets because of the collateral benefit.

To gain some intuition about the conditions under which the borrowing constraint binds, we use equation (19.6). This condition tells us that, keeping everything else constant, a lower cost of debt, $\tilde{R}$, is associated with a higher value of the multiplier, μ. A lower cost of debt makes borrowing more attractive and, therefore, the value of relaxing the borrowing constraint is higher. So, in general, the borrowing constraint is more likely to bind when the effective cost of borrowing, $\tilde{R}$, is lower than the inverse of the expected discount factor m'. The reason we use the term 'likely' is because the effective discount factor for the firm, $m'\varphi'(d)/\varphi'(d')$, depends also on $\varphi'(d)/\varphi'(d')$. This captures the dividend smoothing motive of the firm, similar to consumption smoothing for households which is captured by the stochastic discount factor $m' = \beta u_1(C', L')/u_1(C, L)$.

[7] See Appendix 19.A for the derivation.

19.4.2 Optimality conditions for households

Households maximize their expected lifetime utility by choosing consumption, c; labor supply, ℓ; purchase of bonds from firms, b'; and firms' shares, a'. The households' optimization problem can be written recursively as

$$V(S;a,b) = \max_{c,\ell,a',b'} \left\{ u(c,\ell) + \beta \mathbb{E} V(S';a',b') \right\} \tag{19.7}$$

subject to

$$(d+q)a + b + w\ell = c + qa' + \frac{b'}{R} + T.$$

We denote with q the price of one share which pays the dividend d. R is the gross interest rate and the household lends b'/R today and gets repaid b' in the next period. Notice that $R > \tilde{R}$ since $\tilde{R} = R/(1+\tau)$ and τ is positive. The variable T denotes lump-sum taxes paid by the household to cover the interest subsidies to firms. More specifically, this variable is equal to $T = B'/\tilde{R} - B'/R$, that is, the difference between what firms receive by borrowing, $B'/\tilde{R}$, and what households pay, B'/R. The difference must be covered by taxes. We have used capital letters for the aggregate debt to distinguish it from the debt chosen by an individual household, b'.

From problem (19.7) we derive the first order conditions. Denoting the Lagrange multiplier associated with the budget constraint by γ, the first order conditions lead to[8]

$$u_1(c,\ell)w = -u_2(c,\ell), \tag{19.8}$$

$$\frac{1}{R} = \mathbb{E}\left(\frac{\beta u_1(c',\ell')}{u_1(c,\ell)} \right), \tag{19.9}$$

and

$$q = \mathbb{E}\left[\left(\frac{\beta u_1(c',\ell')}{u_1(c,\ell)} \right) (d'+q') \right]. \tag{19.10}$$

The first equation is the standard optimality condition for the household's supply of labor: it equalizes the utility value of the wage rate earned with one unit of labor (the left-hand side) to the corresponding dis-utility from working an extra unit of time (the right-hand side). The second condition is the optimality condition for bond holdings, which in equilibrium determines the price of the bond (i.e. the interest rate). Finally, the third condition is the optimality condition for share purchases, which in equilibrium determines the price of an equity share. From the last two conditions we can see that the household discounts next period payments by $\beta u_1(c',\ell')/u_1(c,\ell)$, which is also the discount factor for firms.

19.5 Characterization in a two-period version of the model

To gain intuition, it is helpful to consider a simplified version of the model with only two periods where time runs from date 0 to date 1. We specify the utility and production

[8] See Appendix 19.A.2 for a derivation.

functions as

$$u(C_t, L_t) = \ln(C_t - L_t^{\nu})$$

and

$$F(z_t, K_t, L_t) = z_t K_t^{\alpha} L_t^{1-\alpha}.$$

An additional simplification is to abstract from uncertainty. Therefore, in period 0, agents can predict what will happen in period 1. Because time ends after period 1, we impose the terminal conditions $K_2 = 0$ and $B_2 = 0$. This says that all resources will be consumed and the debt must be fully repaid in period 1. We then solve the model backward: we first solve for the equilibrium allocation in period 1 and then we solve for the allocation in period 0. For the rest of this section we indicate variables in period 0 without a subscript and variables in period 1 with a prime superscript.

Terminal period, $t = 1$. Given the state variables in the terminal period—z', K' and B'—we find the supply of labor by solving condition (19.8) after replacing the wage rate w with the marginal product of labor (condition (19.4)). Using the specified functional forms for utility and production, the condition becomes

$$(1-\alpha) z' \left(\frac{K'}{L'}\right)^{\alpha} = \nu L'^{\nu-1}.$$

Solving for labor we obtain

$$L' = \left(\frac{1-\alpha}{\nu}\right)^{\frac{1}{\alpha+\nu-1}} z'^{\frac{1}{\alpha+\nu-1}} K'^{\frac{\alpha}{\alpha+\nu-1}}.$$

Thus, labor increases in TFP, z', and capital, K', as they both raise the productivity of labor. Plugging L' into the production function we obtain

$$Y' = \left(\frac{1-\alpha}{\nu}\right)^{\frac{1-\alpha}{\alpha+\nu-1}} z'^{\frac{\nu}{\alpha+\nu-1}} K'^{\frac{\nu\alpha}{\alpha+\nu-1}},$$

which also increases in z' and K'. In the terminal period all resources are consumed. Therefore, consumption is equal to production, Y', plus non-depreciated capital, $(1-\delta)K'$,

$$C' = (1-\delta)K' + \left(\frac{1-\alpha}{\nu}\right)^{\frac{1-\alpha}{\alpha+\nu-1}} z'^{\frac{\nu}{\alpha+\nu-1}} K'^{\frac{\nu\alpha}{\alpha+\nu-1}}.$$

Finally, the argument of the utility function can be expressed as a function of z' and K', that is,

$$C' - L'^{\nu} = g(z', K').$$

The function $g(z', K')$ is increasing in z' and K', at least in the relevant range, as we can verify by replacing C' and L' with the expressions derived above. This is intuitive because a more productive economy, either because of a higher value of z' or K', will produce more and allows for higher consumption net of the dis-utility from working. Notice that we are making a simplifying assumption: in the terminal period the firm does not incur any cost in the payment of dividends if they deviate from ϕ.

Initial period, $t = 0$. The economy starts the initial period 0 with states z, K and B. Firms choose the input of labor, L, the new stock of capital, K', and the new debt, B'. Households choose the supply of labor and their holdings of bonds. These decisions are determined by the same first order conditions derived from the infinite horizon model.

The input of labor and output are given by the same expressions derived for period 1, that is,

$$L = \left(\frac{1-\alpha}{\nu}\right)^{\frac{1}{\alpha+\nu-1}} z^{\frac{1}{\alpha+\nu-1}} K^{\frac{\alpha}{\alpha+\nu-1}},$$

and

$$Y = \left(\frac{1-\alpha}{\nu}\right)^{\frac{1-\alpha}{\alpha+\nu-1}} z^{\frac{\nu}{\alpha+\nu-1}} K^{\frac{\nu\alpha}{\alpha+\nu-1}}.$$

We focus on the equilibrium in which the borrowing constraint is binding. This would be the case if the initial debt B is sufficiently large. Then, solving for the equilibrium in period 0 entails finding the values of K', B', C, D, R and μ. Having six unknowns, we need six conditions.

The first is the optimal condition for the accumulation of capital, equation (19.20). Using the functional forms specified for the two-period model, the condition can be written as

$$(1-\mu\xi)p = \beta\left(\frac{C-L^{\nu}}{g(z',K')}\right)\Big[1-\delta+F_k(z',K',L')\Big]\Big[1+2\kappa\Big(D-\phi\Big)\Big], \tag{19.11}$$

with the marginal product of capital given by

$$F_k(z',K',L') = \alpha\left(\frac{1-\alpha}{\nu}\right)^{\frac{1-\alpha}{\alpha+\nu-1}} z'^{\frac{\nu}{\alpha+\nu-1}} K'^{\frac{\nu\alpha}{\alpha+\nu-1}-1}.$$

Since we assumed that the borrowing constraint is binding and $\mu > 0$, we can use the borrowing constraint as the second condition,

$$\frac{(1+\tau)B'}{R} = \xi p K'. \tag{19.12}$$

The budget constraint for the firm provides the third condition

$$\alpha F(z,K,L) + \frac{(1+\tau)B'}{R} = B + p\Big[K'-(1-\delta)K\Big] + D + \kappa\Big(D-\phi\Big)^2. \tag{19.13}$$

Next we can use the households' budget constraint,

$$D + B + (1-\alpha)F(z,K,L) = C + \frac{B'}{R} + T, \tag{19.14}$$

where taxes T are equal to $\tau B'/R$.

The fifth and sixth conditions are given by the first order conditions for the issuance of new debt by firms and the purchase of the debt by households:

$$\frac{(1-\mu)(1+\tau)}{R} = \beta\left(\frac{C-L^{\nu}}{g(z',K')}\right)\Big[1+2\kappa\Big(D-\phi\Big)\Big] \tag{19.15}$$

and

$$\frac{1}{R} = \beta \left(\frac{C - L^{\nu}}{g(z', K')} \right). \tag{19.16}$$

Equations (19.11) through (19.16) allow us to solve for K', B', C, D, R and μ under the assumption that the borrowing constraint is binding. This is the equilibrium where financial frictions are more relevant. However, if the borrowing constraint is not binding, to solve for the equilibrium we simply replace the borrowing constraint with $\mu = 0$. This would be the case if B is very small or even negative.

In the next subsections we will use the six conditions (19.11)-(19.16) to characterize the responses of the economy to shocks and to show how the responses are affected by financial frictions.

19.5.1 Financial frictions and propagation of shocks

We start considering a persistent, positive productivity shock, that is an increase in z and z'. The implications are summarized in the following property.

Property 19.5.1 *The response of investment to the productivity shock is positive. Current and future output will also increase. However, the responses of investment and output will be smaller if the borrowing constraint is binding.*

The fact that the macroeconomic responses could be smaller when the borrowing limit is binding implies that financial frictions could dampen the macroeconomic response to productivity shocks rather than amplifying it. This may be surprising but it will become clear once we walk through the equilibrium conditions.

A higher productivity in the next period increases the incentive to invest so that firms would like to increase K'. This can be seen in equation (19.11): keeping everything else constant, an increase in z' raises the marginal productivity of capital $F_k(z', K', L')$. To re-establish equality, K' must rise. Of course, other variables will also change, so that the overall change is more complex. But what we describe here captures the main channel shaping the response to a productivity shock.

The increase in capital needs to be financed in some way by firms. When the firm chooses a higher value of K', the right-hand side of the budget constraint—equation (19.13)—increases. Thus, something on the left-hand side has to increase or something on the right-hand side has to decrease.

The increase in current productivity raises profits, the term $\alpha F(z, K, L)$. This provides some extra funds that can be used to finance investment. Also, a higher value of K' allows the firm to raise more debt as shown in equation (19.12). This increases $(1 + \tau)B'/R$ on the left-hand side of the budget constraint. However, provided that $\xi < 1$, which is typically the case, the increase in borrowing and the extra profits will not be sufficient to fund the desired increase in capital. Therefore, in order for the firm to fund the higher investment, it has to reduce dividends, D.

Now consider the implications from paying lower dividends. Looking at equation (19.11), a lower value of D reduces the last term and discourages capital investment. The intuition is that, reducing dividends has a cost for the firm. If paying lower dividends is necessary

to fund investment, the effective cost of investing will be higher and, as a result, the firm invests less than in absence of this cost.

How should we interpret this in the real world? Due to financial frictions, large investments need to be funded at least in part with equity. But equity is more expensive than other sources of funds, and it becomes more expensive the higher the amount of equity that needs to be raised. This implies that the need to fund investment with equity increases the cost of capital for the firm and discourages investment.

We have shown that, with financial frictions, the response of investment and output could be smaller than in absence of frictions. If that is the case, then financial frictions would dampen the macroeconomic response to shocks rather than amplifying it. The next question is whether there are other mechanisms, besides what we have described here, that could amplify the response of the economy.

19.5.2 The price of capital

An important branch of the macro-finance literature, including the seminal contributions of Bernanke and Gertler (1989) and Kiyotaki and Moore (1997), makes the price of capital endogenous so that p also changes in response to the shock. This is important because the ability to borrow depends on the market value of capital through the borrowing constraint,

$$\frac{(1+\tau)B'}{R} \leq \xi p K'.$$

We could think that the price of capital p increases when the economy experiences a boom, especially if persistent, because an economic expansion is typically associated with a higher demand for capital. An increase in the price of capital raises the net worth of firms because the capital they own is now more valuable. Furthermore, since part of the capital is funded with debt, the value of net worth increases proportionally more than the increase in the price of capital. Higher net worth, then, provides firms with more down payment to fund the acquisition of capital for the next period. We summarize the impact of a change in the price of capital in the following property.

Property 19.5.2 *If the borrowing constraint is binding and the firm starts with positive debt $B > 0$, a positive price change (higher p) has a positive impact on capital accumulation. This increases next period production directly, through a higher input of capital, and indirectly through higher employment.*

To illustrate this idea, consider the net worth of the firm at the beginning of the period. This is the value of capital minus the debt, that is, $pK - B$. Provided that $B > 0$, a 1 percent increase in p generates an increase in net worth that is bigger than 1 percent.

Let's consider now the firm's budget constraint (19.13) and use equation (19.12) to eliminate the new value of debt. We can then re-arrange the budget constraint as

$$K' = \left(\frac{1}{1-\xi}\right)\left((1-\delta)K + \frac{\alpha F(z,K,L) - B - D - \kappa\Big(D-\phi\Big)^2}{p}\right).$$

The term $\alpha F(z,K,L) - B - D - \kappa\Big(D - \phi\Big)^2$ represents the profits net of outstanding liability and dividend payments. Typically, this term is negative because the debt (which is a stock) is bigger than profits (which is a flow). Therefore, an increase in p raises the second term in bracket and results in a higher value of K'. The first term in parentheses on the right-hand side acts as a multiplier, and depends positively on ξ, that is, the fraction of capital that can be used as collateral. Intuitively, a larger ξ allows firms to be more leveraged and, as a result, a change in the price of capital has a larger impact on the firm's net worth. Higher net worth allows the firm to fund more capital.

In summary, an increase in the price of capital generates a proportionally higher increase in the net worth of leveraged firms. The increase in net worth relaxes the borrowing constraint and allows for more investment. This is the central amplification mechanism that operates in Bernanke and Gertler (1989) and Kiyotaki and Moore (1997), as well as in many other models proposed in the macro-finance literature. The basic mechanism illustrated here is also present in the financial accelerator model of Bernanke et al. (1999) and Brunnermeier and Sannikov (2014).

19.5.3 Financial shocks

As described in the previous subsection, the amplification of real shocks could be induced by the impact of the shock on the price of capital. This increases the net worth of borrowers which in turn relaxes the borrowing constraint. In some cases, however, changes in the tightness of the borrowing constraint could be driven by forces that originate directly in the financial sector. This would be captured in our model by a change in the parameter of the borrowing constraint ξ.

There are many mechanisms or channels that could be captured, in reduced form, by a change in ξ. Financial innovations such as securitization could enhance the collateral use of certain assets. This allows borrowers to fund a larger share of investment with debt. It could be balance-sheet difficulties experienced by banks that force them to change the credit standards for the approval of loan applications. Figure 19.2 has shown that bank credit standards are very cyclical in the data. Based on these considerations, some studies in the macro-finance literature assumed that ξ follows a stochastic process and referred to changes in ξ as 'financial shocks' (Jermann and Quadrini, 2012). The goal of this subsection is to understand the macroeconomic implications of these shocks in the context of the simplified two-period model. We summarize the implications as follows.

Property 19.5.3 *If the borrowing constraint is binding, a negative financial shock (lower ξ) has a negative impact on capital accumulation. This reduces next period production directly, through a lower input of capital, and indirectly through lower employment.*

If the borrowing constraint (19.12) is binding, and keeping other things unchanged, a drop in ξ causes a drop in B'. Looking now at the budget constraint of the firm, equation (19.13), we see that the drop in B' reduces the left-hand side of the budget constraint and, therefore, the right-hand side must also fall. This can be done either by paying fewer dividends D, or by cutting investment, that is, choosing a lower value of K'.

Consider first what happens if the firm reduces dividends. From the first order condition that characterizes the optimal borrowing of firms, condition (19.15), we can see that the multiplier μ is likely to increase. Intuitively, since the borrowing constraint is tighter when ξ is lower, the value of relaxing the constraint increases. Remember that the Lagrange multiplier μ represents the value of relaxing the borrowing constraint.

We now turn our attention to the equilibrium condition for investment. A lower payment of dividends D reduces the right-hand side of equation (19.11). The intuition is that, when the firm needs to fund investment by lowering dividends, which is costly, the marginal product of capital must rise. This can be accomplished by reducing K'. We also notice, however, that the left-hand side of condition (19.11) also changes as μ increases (as we observed above) and ξ decreases (which is the shock). Keeping ξ unchanged, the decline in the left-hand side induced by the lower value of μ is smaller than the decline in the right-hand side induced by a lower D. This is because μ is multiplied by ξ, which is typically smaller than 1. At the same time, the value of ξ is now smaller which, for a given μ, increases the left-hand side of condition (19.11). The intuition here is that capital is not only an input of production but it is also a collateral. Using capital as a collateral is valuable because it allows the firm to borrow at a smaller cost than the cost of equity. Therefore, when the borrowing constraint becomes tighter (higher μ), the fraction of capital ξ that can be used as a collateral becomes more valuable, reducing the effective cost of capital. This, however, is counterbalanced by the fact that ξ is now lower.

To summarize, a reduction in the value of ξ has two effects. It reduces the right-hand side of condition (19.11) and it has an ambiguous effect on the left-hand side of (19.11). However, even if the left-hand side drops in value, the drop is smaller than in the right-hand side. Then, in order to re-establish the equality between the left and right sides of equation (19.11), the marginal product of capital $F_k(z', K', L')$ must rise, which requires a lower value of K'. This shows that a decline in ξ (negative financial shock) has a negative impact on investment.

As far as current employment is concerned, we observe that ξ does not affect current employment (see above derivation of L). However, it impacts indirectly next period employment by affecting capital accumulation. From the expression that defines L' derived above, we can see that labor increases in K' and, therefore, lower investment affects adversely employment in the next period.

The effects described here require that the borrowing constraint is binding or at least occasionally binding. This brings us back to the relevance of Assumptions 19.4.1 and 19.4.2: financial frictions create an investment wedge (that is, it distorts investment) and a negative financial shock makes that wedge bigger.

19.5.4 Asymmetric responses

An interesting feature of the model with financial shocks is the potential asymmetry with which the economy responds to positive and negative financial shocks. Suppose that we start from an equilibrium in which the borrowing constraint (19.12) is not binding. This could arise, for example, if the initial capital K is high or the initial debt B is low. Starting from this equilibrium, the economy is hit by a positive credit shock, that is, an increase in ξ. Since firms were not constrained before the shock, the fact that now they can take more

debt does not affect their borrowing: if they were not borrowing up to the limit before, that is, they were optimally borrowing less than what was available to them, there is no reason to change that level of borrowing now that the borrowing capacity has increased. This also means that the shock has not consequences for the economy.

Property 19.5.4 *If the borrowing constraint is not initially binding, then an increase in ξ has not effect on investment, output and employment (current and future).*

Now consider a negative financial shock, that is, a decline in the value of ξ. Provided that the decline in ξ is sufficiently large, the firm will no longer be able to borrow the same amount. In other words, condition (19.12) will no longer be satisfied if the firm chooses the same policies in absence of the change in ξ. The change in policies then will have the effects described in the previous subsection, leading to a contraction in investment and a macroeconomic contraction in the next period. This is more likely to arise if the initial capital K is low and the initial debt B is high.

Property 19.5.5 *If the borrowing constraint is not initially binding, then a sufficiently large decrease in ξ reduces investment, future output and future employment.*

The asymmetric features of financial shocks is one of the reasons they have been used to understand the dynamics of financial booms and busts: while financial booms tend to be gradual and long lasting in the data, financial busts tend to be more sudden with sizable macroeconomic implications. The asymmetry depends on the financial structure of the firm when the shock hits. As argued above, the asymmetry is a direct consequences of binding borrowing constraints which in turn depend on the initial capital K and debt B. But what determines the initial capital and debt? What values are more relevant?

Unfortunately the two-period model is silent about the initial states K and B since they are exogenous. It is in this respect that the infinite horizon model provides additional insights since K and B are endogenous. In particular, through simulation, we can derive the invariant distributions of the states, including K and B. This informs us about the likelihood that certain states emerge in equilibrium and allows us to determine whether binding constraints are common or uncommon. It also informs us about the history that could lead the economy to states where binding constraints become likely. We will show this numerically in the last section of this chapter.

19.5.5 Financial frictions and the labor wedge

So far we have illustrated a particular mechanism through which financial frictions could impact the macro-economy. That mechanism operates through the optimality condition for investment. Financial frictions do not distort directly the optimality condition for labor. In fact, condition (19.4) is exactly the same as in the standard neoclassical model. It affects labor only indirectly through capital accumulation. An implication is that, while the amplification mechanism created by financial frictions could be important for the dynamics of investment, it could be somewhat negligible for the dynamics of labor. Considering that labor contributes significantly to the business cycle dynamics, this would make financial frictions less relevant for understanding the business cycle.

This point can be made more precise starting from the standard Cobb-Douglas production function that, in logarithmic form, can be written as

$$\ln Y_t = \ln z_t + \alpha \ln K_t + (1-\alpha) \ln L_t.$$

After calibrating the capital income share α, we can use data on GDP, capital and labor to measure Y_t, K_t, and L_t. We can then construct a measure of productivity using the above equation, that is, $\ln z_t = \ln Y_t - \alpha \ln K_t - (1-\alpha) \ln L_t$. This is the Solow residual approach used to construct measures of total factor productivity (TFP). We can calculate the contribution of TFP, capital and labor to GDP fluctuations. Setting $\alpha = 0.36$, and using annual data from 1950 to 2019 for the U.S. economy, we find that cyclical movements in labor contribute about 60 percent to the standard deviation of GDP. The contribution of capital, instead, is less than 10 percent. The remaining contribution can be attributed to cyclical fluctuations in TFP.

There are two reasons why movements in the stock of capital have a small contribution to output fluctuations. First, even if capital expenditures are very volatile, they are only a small fraction of the stock of capital. This implies that the stock of capital does not move much over the business cycle even if investment does. Second, movements in the stock of capital are multiplied by the share α, which is smaller than 0.5. Movements in labor, instead, are multiplied by $1-\alpha$ which is bigger than 0.5. Therefore, if financial frictions play a more relevant role for aggregate output fluctuations, they must have a more direct impact on labor. One way to do that is with the introduction of working capital.

Suppose that the payment of wages needs to be made before firms receive the revenues from their sales. This is a feature of the production cycle where costs and revenues are not perfectly synchronized. Because of this, firms need to carry liquid funds from the previous period or borrow additional funds to pay for the wages. Carrying liquid funds from the previous period or borrowing in the current period have similar implications. However, the latter is more convenient in terms of notation. We thus assume that firms borrow at the beginning of the period to make advanced payments of wages. This type of borrowing is in addition to the intertemporal debt already introduced in the model. Since the debt raised to pay wages is repaid within the period, there are no interest payments.

With the working capital extension we have two types of borrowing: the intra-period debt, which is equal to the wage bill, $w\ell$, and the inter-temporal debt b'/R. The total debt, sum of the two types of borrowing, is subject to the same limit as before, that is,

$$w\ell + \frac{b'}{R} \leq \xi p k'. \tag{19.17}$$

The problem solved by the firm is still given by (19.3) but with the borrowing constraint taking the form specified in (19.17). The optimality conditions for capital and intertemporal borrowing are (19.5) and (19.6), which are the same as before. The first order condition for labor, however, changes to

$$F_3(z, k, \ell) = (1+\mu)w. \tag{19.18}$$

This shows that the marginal product of labor is equalized to the wage rate only if the multiplier μ is zero, that is, if the borrowing constraint is not binding. However, if the

borrowing constraint is binding, the input of labor will be distorted, and the distortion increases in the tightness of the borrowing constraint, which is captured by the multiplier μ.

This introduces an interesting mechanism through which financial frictions could have important implications for aggregate economic activity: tighter financial conditions will be reflected in higher values of the multiplier μ which, for a given wage rate, reduce the demand for labor.

Productivity and financial shocks affect the macro-economy in a similar way as described earlier, but now they impact the demand for labor directly. We show this again with the two period model where, for simplicity, we assume that wages are paid in advance only in period 0. Thus, the working capital constraint does not apply in the terminal period 1. This also implies that the equilibrium in period 1, given the states z', K' and B', is equivalent to the one characterized earlier. We state the main property as follows.

Property 19.5.6 *Given productivity z and capital K, employment declines with the tightness of the borrowing constraint, which is captured by the multiplier μ.*

In period 0, equilibrium labor is found by solving condition (19.19), adjusted for the presence of working capital. Using the specific functional forms for utility and production, the condition is

$$(1-\alpha)zK^{\alpha}L^{-\alpha} = (1+\mu)\nu L^{\nu-1}.$$

Solving for labor we obtain

$$L = \left[\frac{1-\alpha}{(1+\mu)\nu}\right]^{\frac{1}{\alpha+\nu-1}} z^{\frac{1}{\alpha+\nu-1}} K^{\frac{\alpha}{\alpha+\nu-1}}.$$

Labor continues to increase in productivity, z, and capital, K. In addition, it now depends negatively on the multiplier μ. This is because, as the borrowing constraint becomes tighter and the value of μ increases, firms demand less labor and, in equilibrium, employment is lower.

We could repeat the same analysis conducted earlier with the two-period model and obtain similar results. The most important addition is that a decline in the price of capital p and/or a decline in ξ generates a macroeconomic contraction also in period 0 since it affects directly the demand for labor. This is in addition the to impact on investment and on next period production as described earlier.

We conclude this subsection by pointing out that the need for working capital does not derive only from the advanced payment of wages. It could also derive from the advance payment of intermediate inputs. The production function typically used in the standard neoclassical growth model abstracts from the need of intermediate inputs. Instead, it focuses on the reduced form where final goods are produced only with the inputs of capital and labor. In reality, though, any type of production requires intermediate goods that are produced by other firms and may need financing. With a more complex production process, tighter financial conditions may cause a macroeconomic contraction by distorting the input of intermediate goods.

19.6 General model and the neoclassical growth model

The equilibrium in the infinite horizon model can be characterized by combining the first order conditions of households and firms together with market clearing. Replacing the wage rate in equation (19.8) using equation (19.4) we obtain

$$F_3(z, K, L) = -\frac{u_2(C, L)}{u_1(C, L)}. \tag{19.19}$$

The marginal product of labor is always equalized to the marginal rate of substitution between consumption and leisure, which is the same condition for the standard neoclassical growth model.

Equation (19.5) is the optimality condition for the accumulation of capital. We can rewrite it as

$$(1 - \mu\xi)p = \mathbb{E}\left(\beta\frac{u_1(C', L')}{u_1(C, L)}\right)\left(\frac{\varphi'(D)}{\varphi'(D')}\right)\Big[(1 - \delta)p' + F_2(z', K', L')\Big]. \tag{19.20}$$

Compared to the standard neoclassical growth model, we observe two important differences. First, the effective cost of capital (the left-hand side) is not the price p but $(1 - \mu\xi)p$ reflecting the collateral value of capital discussed earlier.

The second difference with the standard neoclassical growth model is that the return from capital received in the next period (marginal product plus resale value) is discounted at a different rate. The effective discount factor used by firms includes the term $\varphi'(D)/\varphi'(D')$, which is absent in the standard model. Obviously, this term will disappear if we relax Assumption 19.4.3 and there is no cost in the payment of dividends. If we also relax Assumption 19.4.2, then $\mu = 0$ and the equilibrium condition for capital would be exactly identical to the standard neoclassical growth model.[9]

We now turn to the equilibrium in the bond market (the debt issued by firms and purchased by households). Using $\tilde{R} = R/(1 + \tau)$, equation (19.6) can be rewritten as

$$\frac{(1 - \mu)(1 + \tau)}{R} = \mathbb{E}\left(\beta\frac{u_1(C', L')}{u_1(C, L)}\right)\left(\frac{\varphi'(D)}{\varphi'(D')}\right), \tag{19.21}$$

while the first order condition for the households' choice of bonds, equation (19.6), is

$$\frac{1}{R} = \mathbb{E}\left(\beta\frac{u_1(C', L')}{u_1(C, L)}\right). \tag{19.22}$$

If we relax Assumption 19.4.3 by imposing $\kappa = 0$ (no cost in deviating from the dividend target), conditions (19.21) and (19.22) imply $(1 - \mu)(1 + \tau) = 1$. Therefore, a sufficient condition to have a binding borrowing constraint (so that $\mu > 0$) is that $\tau > 0$. This is Assumption 19.4.1. However, if $\kappa > 0$, which implies $\varphi(D) \neq D$, the borrowing constraint may not be always binding. For example, if in expectation $\varphi'(D)/\varphi'(D')$ is sufficiently bigger

[9] For consistency, however, we cannot relax Assumption 19.4.2 while keeping Assumption 19.4.1. Otherwise the firm would borrow an infinite amount of debt. Thus, the relaxation of Assumption 19.4.2 is done jointly with the relaxation of Assumption 19.4.1.

than 1, then condition (19.21) cannot be satisfied with $\mu > 0$. This could arise if the firm would like to pay more dividends now relatively to those paid in the next period.

Another contingency in which this may arise is when, in expectation, the term $u_1(C', L')/u_1(C, L)$ is higher. For example, when future households' consumption is lower than current consumption. The property that the borrowing constraint could be occasionally binding has been explored in the financial crises literature: a crisis could be a contingency in which the borrowing constraint becomes binding. We will come back to this point in the next section when we conduct a numerical analysis.

To summarize, we have shown that Assumptions 19.4.1, 19.4.2, 19.4.3 introduce some wedges in the equilibrium condition for the accumulation of capital that distort allocations. In absence of these three assumptions we would revert to the standard neoclassical model as stated formally in the following proposition.

Proposition 19.1 *If $\tau = 0$ and $\kappa = 0$, the equilibrium debt chosen by firms is indeterminate and the real allocation is the same as in the standard neoclassical growth model.*

Proof. The proof is obtained by comparing the equilibrium first order conditions after setting $\tau = 0$ and $\kappa = 0$. We have already shown that with $\tau = 0$ and $\kappa = 0$, the multiplier associated with the borrowing constraint is $\mu = 0$. Conditions (19.20) and (19.21) then become

$$p = \beta \mathbb{E} \left(\frac{u_1(C', L')}{u_1(C, L)} \right) \left[(1 - \delta) p' + F_2(z', K', L') \right]$$

and

$$\frac{1}{R} = \beta \mathbb{E} \left(\frac{u_1(C', L')}{u_1(C, L)} \right).$$

These two conditions, together with condition (19.19), characterize the equilibrium. Since they are exactly the same conditions as in the neoclassical model, the equilibrium allocation is the same. ■

If $\tau = 0$, the cost of debt for the firm, $\tilde{R}$, is equal to the interest rate, R. If $\kappa = 0$, there is no cost in deviating from the targeted dividend ϕ. This is equivalent to eliminating Assumptions 19.4.1 and 19.4.3. In equilibrium, then, firms are indifferent between funding investment with debt or equity. This is true even if there is a borrowing limit as assumed in Proposition 19.4.2. This is another example of the Modigliani-Miller principle we discussed in Section 19.3.1.

More intuition for Modigliani-Miller principle

Consider a change in financial structure where the firm raises Δ units of funds with debt and uses the funds to pay dividends. By doing so the firm increases its liabilities and reduces its equity. In the next period, then, the firm reduces the payment of dividends to repay the debt inclusive of interests, ΔR. By doing so the firm increases dividend now by Δ and reduces the next period dividends by ΔR. In present value, the next period dividends are worth $\mathbb{E} m' \Delta R$. Therefore, the gain from changing the financial

structure is $\Delta(1 - \mathbb{E}m'R)$. When $\tau = \kappa = 0$, the first order condition (19.6) simplifies to $\mathbb{E}m'R = 1$. This implies that the gain from changing the sources of funding is zero.

The proposition also says that the equilibrium allocation remains the same as in the standard neoclassical model even if investment decisions are made by firms instead of households. Intuitively, this is because firms use the same discount factor as households and, therefore, they make the same decisions as those made by households.

19.7 Numerical analysis

We now use the infinite-horizon model with working capital to show some of the properties numerically. While a two-period model is useful to gain intuition, it necessarily takes the initial states as given. As we observed in the analysis of the two-period model, whether the equilibrium features binding or non-binding borrowing constraints depends on the initial states. So, the two-period model does not offer much guidance about whether the borrowing constraint binds or the severity of financial frictions more generally.

19.7.1 Calibration

Let's first specify the households' utility, which we assume to take the form

$$u(c, \ell) = c^{\nu}(1 - \ell)^{1-\nu}.$$

We calibrate the model at a quarterly frequency and we impose that the price of capital is constant and equal to 1 ($p = 1$). As observed above, many components of the model are similar to the RBC model. Therefore, for the parameters that are common to the two models, we use the values typically used in the RBC literature. These parameters include the discount factor $\beta = 0.983$, the weight of consumption in the utility function $\nu = 0.412$, the share of capital in the production function $\alpha = 0.36$, and the process for productivity z. It is customary to assume that TFP follows a first order Markov process. We do the same here but we assume that the shock takes only three values (a finite-state Markov chain). We calibrate these values together with the transition probability matrix to match the correlation and standard deviation of the empirical series of Solow residuals, as done in the RBC literature. The mean value of z is not important as it acts as a normalization factor.

At this point, we have the values for the standard parameters and we are left with the non-standard parameters. They include the parameters that determines the firm's benefit of debt over equity, τ; the cost of dividend deviations, κ; and those determining the stochastic process for the variable ξ (financial shock). These are parameters that are absent in the RBC model and, therefore, their calibration requires a more detailed description.

We start with τ, the tax benefit of debt for the firm. Since corporations can deduct interest payments when determining corporate tax liabilities, the effective cost of debt is reduced by the corporate tax rate. Assuming a marginal corporate tax rate of 28.1%, we set $\tau = 0.281$.

Let us try now to understand which variables are more likely to be affected by the remaining parameters, starting with the stochastic process for ξ. As for productivity, we

assume that ξ follows a three-state Markov process. The mean value of ξ has a direct effect on the average debt chosen by the firm. Therefore, one empirical moment we can use as a calibration target is the average leverage observed in the data (the ratio of debt B over physical capital K). A second moment we can use is the persistence of debt in the data (autocorrelation) since in the model the persistence of debt will be related to the persistence in ξ. Finally, we can use the volatility of debt in the data (standard deviation) as the third targeted moment: a more volatile ξ will generate a more volatile debt in the model. Thus, the empirical measures of mean, autocorrelation and standard deviation of debt will be the three moments we can use to calibrate the parameters that determine the stochastic process for ξ.

The last parameter we need to calibrate is κ. Higher is the value of κ and lower is the volatility of dividends. Based on this, we can use the empirical volatility of dividends as a calibration target. Since dividends in the model capture not only dividends but more generally payout to shareholders, we can use the volatility of equity payout defined as dividends plus share repurchases minus new equity issuance. The full set of parameters with their calibrated values are reported in Table 19.1.

Table 19.1: Parameter values.

Parameters	Description	Values
β	Discount factor	0.983
ν	Utility parameter	0.412
α	Capital share in production	0.360
τ	Benefit of debt	0.281
κ	Dividends' cost	0.500
p	Price of capital	1.000
z	Productivity and transition probabilities	$\begin{pmatrix} 0.183 \\ 0.185 \\ 0.187 \end{pmatrix}, \begin{bmatrix} 0.900 & 0.075 & 0.025 \\ 0.050 & 0.900 & 0.050 \\ 0.025 & 0.075 & 0.900 \end{bmatrix}$
ξ	Debt limit and transition probabilities	$\begin{pmatrix} 0.450 \\ 0.500 \\ 0.550 \end{pmatrix}, \begin{bmatrix} 0.900 & 0.075 & 0.025 \\ 0.050 & 0.900 & 0.050 \\ 0.025 & 0.075 & 0.900 \end{bmatrix}$

19.7.2 Numerical solution of the model

The model is solved numerically using a global method. We first discretize the state space for K and B/K on a two-dimensional grid and iterate on the optimality conditions using the projection method. The discretization of B/K rather than B is convenient because, given the borrowing limit $B/\tilde{R} \leq \xi K$, the admissible values of B change with K while the admissible values of B/K are independent of K. The detailed computational steps are described in the online appendix. The online appendix also provides the codes to replicate the results shown here.

19.7.3 Simulation exercise

The numerical analysis is based on the following thought experiment. Suppose that the economy experiences a long sequence of productivity and financial shocks $z = 0.185$ and $\xi = 0.500$. These are the middle values of the three possible realizations and correspond to the mean of the two shocks. Even if the long sequence of shocks have the same values, agents do not anticipate them. Thus, in every period, they predict the next period realizations using the conditional probabilities. After the long sequence of shock realizations, the economy converges to an equilibrium that remains constant until different realizations of the two shocks are drawn. Starting from this equilibrium, we assume that the realization of productivity switches from $z = 0.185$ to $z = 0.187$—the highest possible realization of productivity—and stays there for several periods. The draw of the financial shock, instead, remains at $\xi = 0.500$ over the whole simulation. The engineered switch in z is meant to capture a productivity boom and the goal of the exercise is to explore how the various endogenous variables respond to the productivity boom.

To compare the productivity boom to a productivity decline, we repeat the simulation just described but with productivity that switches to $z = 0.183$—the lowest value—and stays there for some periods. Also in this case the draw of the financial shock remains $\xi = 0.500$ over the whole simulation. The simulation captures a productivity downturn. The dynamics of debt and output around the two productivity switches are plotted in Figure 19.3.

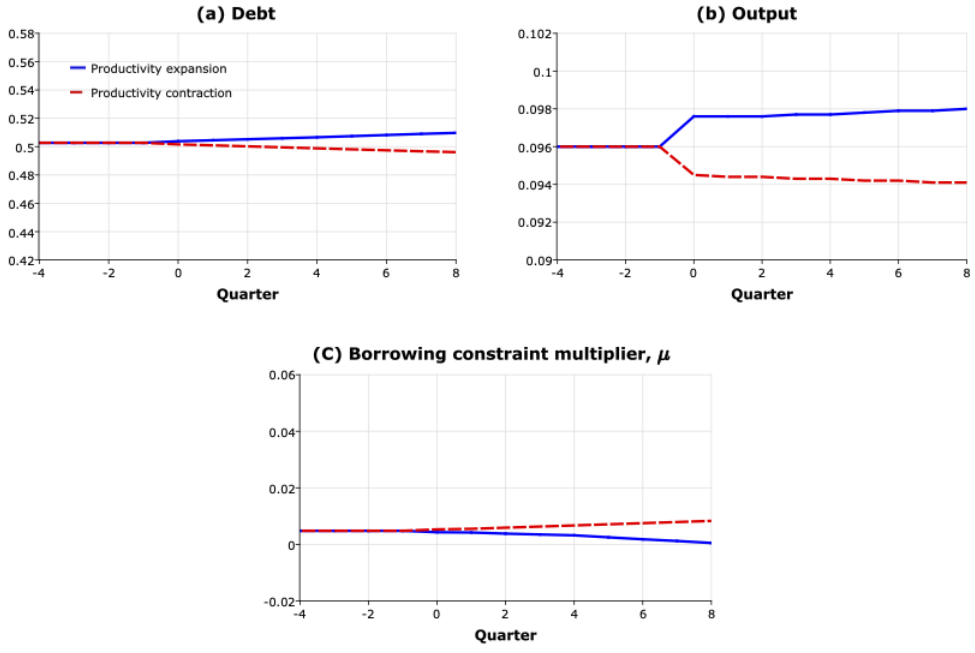


Figure 19.3: Response of debt, output, and borrowing constraint multiplier to a persistent productivity shock.

The switch in productivity arises at quarter zero. We can see that the productivity expansion and contraction lead to the opposite but symmetric dynamics of debt and output. Panel (C) plots the Lagrange multiplier μ, which takes a positive value when the borrowing

limit is binding. As can be seen, μ is always positive meaning that the borrowing constraint is always binding over the whole simulation.

Are the responses of financial variables consistent with the empirical facts outlined by Figures 19.1 and 19.2? The responses of debt and output are both positive, implying positive co-movement as in the data. However, we observe that the response of debt is relatively small compared to the response of output while the data show the opposite pattern. What about interest rate spreads? In the model we do not have interest rate spreads because there is not default. However, to the extent that interest rate spreads capture the tightness of financial constraints, we can compare the dynamics of the multiplier μ to the spreads in the data. As shown in Figure 19.2, spreads are highly counter-cyclical, which is also the case for the responses of μ to productivity shocks. A similar observation can be made regarding loan standards which can be interpreted as indicators of financial tightness. In the model, financial tightness is captured by the multiplier μ. Both loan standards in the data and μ in the model are counter-cyclical. Our simulation assumes that the price of capital p is fixed. However, if the price of capital increases in a productivity boom, it could relax the borrowing constraint and lead to a much lower value of μ. Conceptually this is possible but for this to be important quantitatively, the movements in p must be sizable. It turns out that this is challenging to achieve in models that maintain the basic structure of the RBC framework.

We now turn our attention to financial shocks. We conduct a similar quantitative exercise but with changes in the financial variable ξ. We assume that the economy experiences a draw of a long sequence of productivity and financial shocks $z = 0.185$ and $\xi = 0.500$. At some point, however, the realization of the financial variable increases from $\xi = 0.500$ to $\xi = 0.550$, and stays at the higher level for some periods. The draw of productivity, instead, remains at its mean value $\xi = 0.185$. This corresponds to a credit boom. We then repeat the simulation but this time assuming that the financial variable switches to the lower value $\xi = 0.450$, while productivity remains constant throughout the simulation. The dynamics of debt, output and multiplier are shown in Figure 19.4.

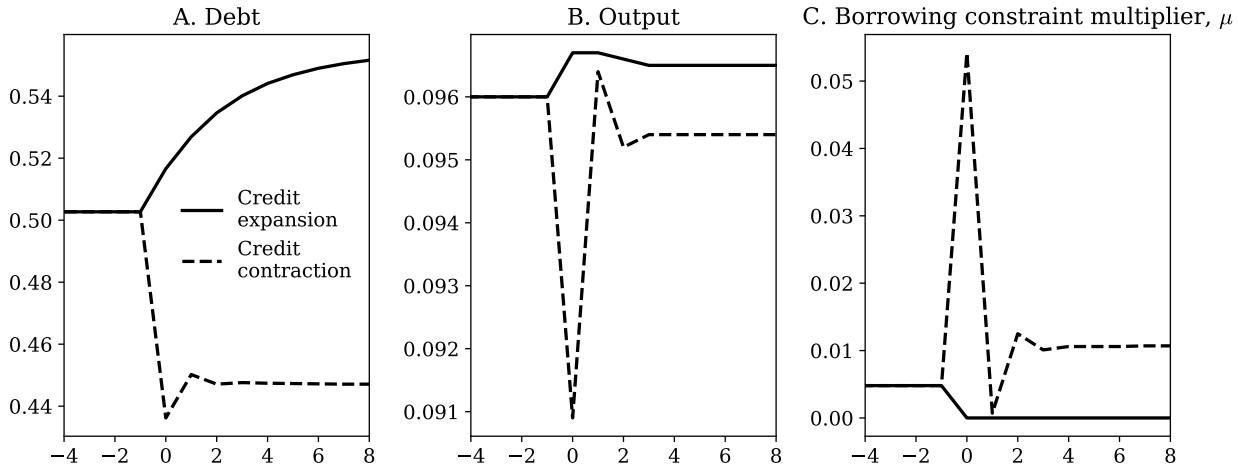


Figure 19.4: Response of debt, output, and borrowing constraint multiplier to a persistent financial shock.

Even though the switch in the financial shock is symmetric (the increase in ξ in the

credit expansion is equal to the decrease in ξ in the credit contraction), the responses of debt and output are asymmetric. A positive financial shock leads to a gradual increase in debt and to a relatively persistent but mild increase in output. On the contrary, a negative financial shock leads to a sharp decline in debt and to a very large contraction in output. The contraction, however, is not very persistent. Panel (C) shows that the borrowing constraint becomes non-binding after the credit expansion, as the multiplier becomes zero.

What is the intuition for the dynamics shown in Figure 19.4? Once ξ reaches the highest value $\xi = 0.550$, there is always the possibility of reversal which would force firms to re-adjust their debt if they borrow up to the limit. But this will be very costly because it requires the firm to either reduce dividends (which is costly because of $\varphi(\cdot)$) or to reduce labor and investment. This creates a precautionary motive that induces firms to borrow less than the limit, even if debt is cheaper than equity.

When the financial shock is negative, the firm is forced to cut borrowing. In order to do so, the firm needs to cut labor and investment (so that less financing is needed) or pay lower dividends (so that the reduction in debt can be compensated by an increase in equity). Both choices are inefficient and the firm chooses the optimal combination that minimizes the inefficiencies.

The qualitative properties of the dynamics induced by financial shocks shown in Figure 19.4 is consistent with the stylized facts outlined in the empirical literature. Reinhart and Rogoff (2009) and Schularick and Taylor (2012), for example, have shown that many episodes of credit booms are not associated with much faster growth in real economic activity. However, when the credit boom experiences a sudden stop, the reversal is often characterized by sharp macroeconomic contractions.

The dynamics displayed by the model are also consistent with the empirical dynamics shown in Figures 19.1 and 19.2. In addition to having pro-cyclical debt, we also have that financial tightness (interest rates spreads and loan standards in the data, and multiplier μ in the model) is counter-cyclical.

Chapter 20

Frictional labor markets

Toshihiko Mukoyama and Ayşegül Şahin

20.1 Introduction

Early real business cycle models have assumed a frictionless labor market. In a frictionless labor market, all firms can find workers, and all workers can find jobs with the equilibrium wage equating labor demand and labor supply. The change in aggregate employment reflects shifts in either the labor demand curve or the labor supply curve.

These models ignore unemployment, that is, the phenomenon that some workers who want to work and look for jobs cannot find jobs. The unemployment rate, defined as the fraction of unemployed workers in the labor force, is an important indicator of the business cycle. The unemployment rate tends to increase when the macroeconomy is in a recession. The elevated unemployment rate is often regarded as one of the most important social costs of recessions. Various government policies, such as unemployment insurance and job training, have been implemented to reduce unemployment and address issues arising from unemployment. To analyze these policies, we need to develop a formal theoretical framework where unemployment arises endogenously.

This chapter introduces labor market frictions in macroeconomic models to analyze unemployment. There are many theories of unemployment in macroeconomics. One simple theory is that there are frictions in wage adjustment. If the wages are too high compared to the level that clears the market, the quantity supplied can exceed the quantity demanded in the labor market. The excess supply of labor can be interpreted as unemployment. Wages can be too high, for example, because of institutional reasons such as minimum wages or unions, or there can be economic reasons. The theory of efficiency wages, for example, postulates that employers want to keep the wage high so that they can induce the workers to exert a high level of effort at work.

In this section, we focus on unemployment arising from search frictions. In the models with search frictions, it takes time, effort, and resources for workers and firms to match with each other. The model describes the element of reality that it takes time for a worker to find a job that is sufficiently good for them, and it takes time for a firm to find a worker who can perform the task that the job requires. In principle, these frictions can exist in many markets (e.g., it may take time to find the kind of chocolate one wants to buy). Even so, we can easily imagine that this type of friction is particularly severe in the labor market because workers

and jobs are heterogeneous in many dimensions. Some search models explicitly deal with the matching of heterogeneous workers and jobs. Some models treat the matching process as a "black box" and use reduced-form functions to formulate it. An example of such a model is the Diamond-Mortensen-Pissarides (DMP) model, described in Section 20.4 below. As we will see, the DMP model has the advantage of being able to fit some of the salient features of the labor market.

20.2 Some labor market facts

Figure 20.1 plots the unemployment rate in the postwar U.S. economy, computed from the Current Population Survey (CPS). In the statistics provided below, the entire U.S. civilian noninstitutional population (16 years old and above) is divided into employment (E), unemployment (U), and not in the labor force (N). The unemployment rate is defined as $U/(E+U)$. In the figure, shaded periods indicate recessions defined by the National Bureau of Economic Research.[1] Unemployment rate is described as the ratio of workers who cannot find a job (although they are searching for a job or on temporary layoff) to the entire labor force. The figure clearly indicates that the unemployment rate increases during recessions. There is no apparent long-run trend in the unemployment rate. Whereas the unemployment rate trended up in the 1970s, it has trended down since then, except for the stark increases in the Great Recession and the COVID-19 pandemic.

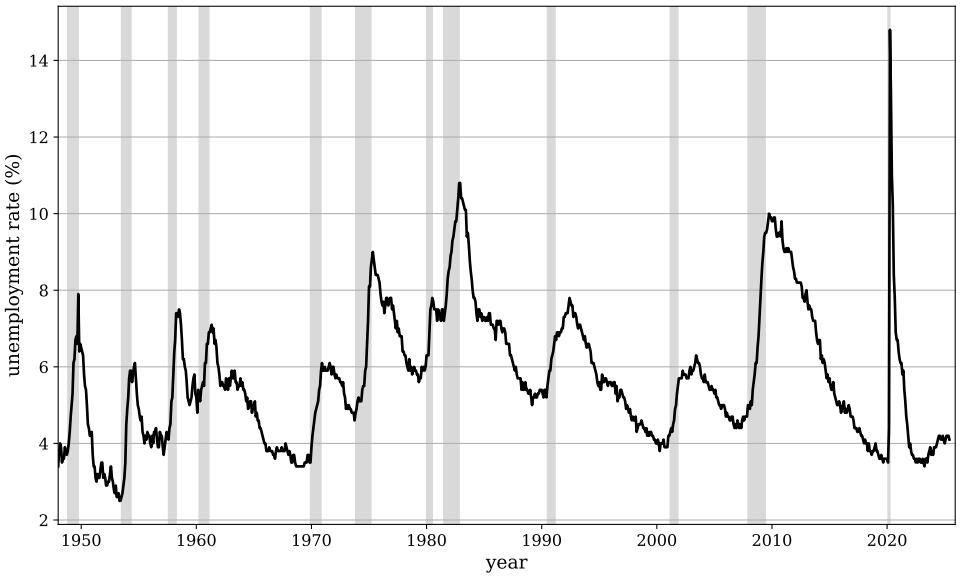


Figure 20.1: Unemployment rate in the United States.

Source: CPS.

Figure 20.2 plots the unemployment rate and the vacancy rate (V represents vacancy) in the United States from December 2000 to May 2022. The unemployment rate is identical to Figure 20.1, and the vacancy rate is computed from the Job Openings and Labor Turnover

[1] See `https://www.nber.org/research/business-cycle-dating`.

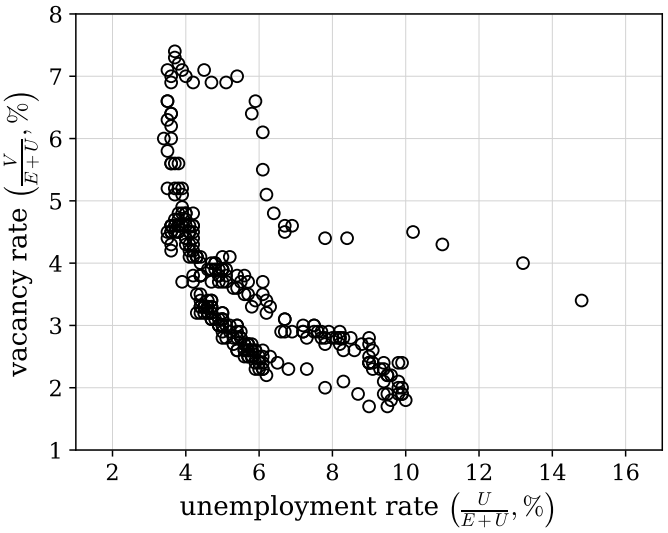


Figure 20.2: Unemployment rate and vacancy rate in the United States.

Source: JOLTS and CPS.

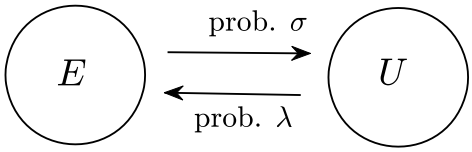


Figure 20.3: The simple model of unemployment.

Survey (JOLTS) and the CPS.[2] Here, we simply point out that vacancy and unemployment coexist in the labor market, indicating that there are trivial amounts of friction in the labor market. We will come back to this figure later on.

20.3 A simple model of unemployment

As a starting point, consider a simple model of unemployment. The model can be simply described by Figure 20.3. Workers are either employed (E) or unemployed (U). We normalize the total labor force to 1, and therefore

$$e_t + u_t = 1,$$

where e_t is employment at period t and u_t is unemployment at period t. We ignore the movements in and out of the labor force. Because the labor force is 1, u_t is the labor force at period t. We assume unemployed workers transit into employment with probability λ and employed workers move into unemployment with probability σ. The probability λ is often called the *job finding probability* and the probability σ is called the *separation probability*.

[2]JOLTS defines the job opening rate as $V/(E+V)$. We transform it to $V/(E+U)$, which is a more relevant object for the theoretical framework below.

Let the unemployment rate at period t be u_t. Then

$$u_{t+1} = (1-\lambda)u_t + \sigma(1-u_t) \tag{20.1}$$

holds, where the first term on the right-hand side is the unemployed workers at period t who remain unemployed in period $t+1$, and the second term is employed workers who lose their jobs between period t and $t+1$. In the steady state where the unemployment rate is constant at $u_{t+1} = u_t = \bar{u}$,

$$\bar{u} = (1-\lambda)\bar{u} + \sigma(1-\bar{u})$$

holds, and therefore

$$\bar{u} = \frac{\sigma}{\lambda+\sigma}. \tag{20.2}$$

The steady-state unemployment rate is decreasing in the job-finding probability λ and increasing in the separation probability σ. Note the steady state, characterized by the expression (20.2), is unique. Moreover, using (20.1) for period t and $t-1$, we obtain

$$u_{t+1} - u_t = (1-\lambda-\sigma)(u_t - u_{t-1}),$$

and because $|1-\lambda-\sigma| \in (0,1)$, the sequence of the unemployment rate is a Cauchy sequence, and thus it converges to the steady-state value. Similarly

$$u_{t+1} - \bar{u} = (1-\lambda-\sigma)(u_t - \bar{u})$$

holds, and we can see that $|1-\lambda-\sigma|$ determines the speed of convergence. If, for example, monthly $\lambda = 0.45$ and $\sigma = 0.034$ (the numbers we use later in the quantitative analysis), $1-\lambda-\sigma$ is almost half, and u_t converges to the steady-state value very quickly. In this case, $\bar{u}$ can be computed as about 7.0%, and if the economy starts from an unemployment rate of 15%, after six months, the unemployment rate goes down to 7.2%.

The job-finding rate λ is also closely linked to the average (or expected) duration of unemployment. As the probability that the unemployment duration is one period is λ, the probability that it is two periods is $(1-\lambda)\lambda$, the probability that it is three periods is $(1-\lambda)^2\lambda$, and so on, the average duration of unemployment D can be computed as:

$$\begin{aligned} D &= \lambda \cdot 1 + (1-\lambda)\lambda \cdot 2 + (1-\lambda)^2\lambda \cdot 3 + \cdots \\ &= \lambda\{[1 + (1-\lambda) + (1-\lambda)^2 + \cdots] + [(1-\lambda) + (1-\lambda)^2 + \cdots] + [(1-\lambda)^2 + \cdots] + \cdots\} \\ &= \lambda\left[\frac{1}{\lambda} + (1-\lambda)\frac{1}{\lambda} + (1-\lambda)^2\frac{1}{\lambda} + \cdots\right] \\ &= \frac{1}{\lambda}. \end{aligned}$$

Thus the average duration of unemployment is inversely related to λ. Note that D can also be derived from a recursive formulation. Let the expected duration of unemployment from the viewpoint of time t as D_t. Then

$$D_t = \lambda \cdot 1 + (1-\lambda)(1 + D_{t+1}),$$

because if the worker finds a job next period (with probability λ), the duration is 1, and if she doesn't, it is one period plus the expected duration from the next period. Here, there is no difference in the expected duration forward at period t and period $t+1$, and thus $D_t = D_{t+1} = D$, and we can solve $D = 1/\lambda$.

20.4 The Diamond-Mortensen-Pissarides (DMP) model

This section introduces a basic search and matching model often referred to as the Diamond-Mortensen-Pissarides (DMP) model. The model in this section is a discrete-time version of Pissarides (1985b).[3] It is called the "search and matching model" because workers and firms have to engage in search activity (in this model, only firms engage in an active search effort, but both the workers and the firms have to wait until they find their counterpart), and the probability of a successful search is governed by a function called the matching function.

The DMP model features an active search by firms in the form of vacancy postings. Vacancy posting is a form of investment: pay the cost now and receive payoffs later. Firms and workers share the surplus (the difference between market production and home production), and the firm can receive a positive profit. In that sense, we also depart from the competitive labor market. Because the firms are the only party that engage in active search activities, this model focuses on the demand side of the labor market. As we will see, this assumption is consistent with the behavior of vacancies in the labor market.

20.4.1 Matching function and the labor market dynamics

We assume that workers are either employed or unemployed. The total population is normalized to 1. The basic structure of the model is similar to Section 20.3, and we can view the model of this section as endogenizing λ of that simple model.

Firms that look for workers post vacancies to search. The number of vacancies is endogenous—the vacancy posting behavior of firms responds to the costs and benefits of hiring workers. Vacancy posting is a (risky) investment for firms: it is costly to post a vacancy, but if firms successfully hire workers, they can enjoy the profit arising from production in the future. Unemployed workers search for firms to work for. The matching process between firms (vacancies) and unemployed workers is summarized by the *matching function*:

$$\mathcal{M}_{t+1} = M(u_t, v_t),$$

where $\mathcal{M}_{t+1}$ is the number of matches created at the beginning of period $t+1$. The function $M(\cdot, \cdot)$ is increasing in both terms and exhibits constant returns to scale. It also satisfies $M(u_t, v_t) \leq u_t$ and $M(u_t, v_t) \leq v_t$. This function is a "black box" that summarizes the complex process of firms' recruiting activities. In particular, workers and firms are heterogeneous, and it is not an easy task for a firm to find a suitable worker for its position. Different firms do not coordinate their recruiting efforts, and they may go after the same person even when there are other people available. The interpretation of the matching function can vary across different models, but in the basic DMP model, the "black box" is interpreted as incorporating all difficulties firms face when recruiting workers.

We assume that the search is random, that is, all vacancies have the same chance of finding workers, and all workers have the same chance of meeting the vacancies. Thus the probability of a worker meeting a firm is

$$\frac{M(u_t, v_t)}{u_t} = M\left(1, \frac{v_t}{u_t}\right) = M(1, \theta_t),$$

[3] The textbook Pissarides (2000) explores various versions of the DMP model in continuous time.

where θ_t is defined as $\theta_t \equiv v_t/u_t$ and often referred to as the labor market tightness. Let us denote

$$\lambda_w(\theta_t) \equiv M(1, \theta_t). \tag{20.3}$$

This $\lambda_w(\theta_t)$ corresponds to λ in Section 20.3. Note that $\lambda_w(\cdot)$ is increasing in θ_t from our assumptions about the matching function. The probability of a vacancy meeting a worker is

$$\frac{M(u_t, v_t)}{v_t} = M\left(\frac{u_t}{v_t}, 1\right) = M\left(\frac{1}{\theta_t}, 1\right).$$

Let us define

$$\lambda_f(\theta_t) \equiv M\left(\frac{1}{\theta_t}, 1\right).$$

Note that

$$\lambda_w(\theta_t) = \theta_t \lambda_f(\theta_t)$$

holds.

It turns out that, when $z > b$, all firms and workers accept all matches once they meet. Thus $\lambda_w(\theta_t)$ represents the job-finding probability of an unemployed worker. It also represents the probability that a worker transitions from unemployment to employment, and unlike in Section 20.3, it depends on the labor market tightness. In this section (similarly to Section 20.3), we assume that matches resolve with probability $\sigma \in (0, 1)$.

Therefore, the dynamics of unemployment is governed by

$$u_{t+1} = (1 - \lambda_w(\theta_t))u_t + \sigma(1 - u_t). \tag{20.4}$$

The first term on the right-hand side is the unemployed workers at period t who stay unemployed at period $t + 1$. The second term is the employed workers (note that $e_t = 1 - u_t$) who separate from their jobs between t and $t + 1$.

When the vacancy is constant at v, it is straightforward to show that there exists a unique steady state value of u_t (call it $\bar{u}$). To see this, set $u_{t+1} = u_t = \bar{u}$ in (20.4) and obtain

$$\bar{u} = \frac{\sigma}{\lambda_w(v/\bar{u}) + \sigma}. \tag{20.5}$$

This equation can be rewritten as, using (20.3),

$$M(v, \bar{u}) + \sigma\bar{u} = \sigma. \tag{20.6}$$

Because the right-hand side is constant and the left-hand side is increasing in $\bar{u}$, the solution of $\bar{u}$ in (20.5) is unique. Further note that (20.6) describes a negative relationship between v and $\bar{u}$ when σ is kept constant.

Now let us go back to Figure 20.2. Figure 20.4 connects the data points of Figure 20.2. It is clear that there is a negative relationship between the unemployment rate and the vacancy rate. This regularity is often called the *Beveridge curve*. The Beveridge curve relationship is consistent with the equation (20.6), and therefore provides support for this component of the DMP model. For this reason, equation (20.6) is often referred to as the Beveridge curve relationship. Moreover, the strong procyclical movement of vacancy indicates that the firm's recruiting activities (i.e., the labor demand movements) play an important role in driving the cyclical movement of the unemployment rate.

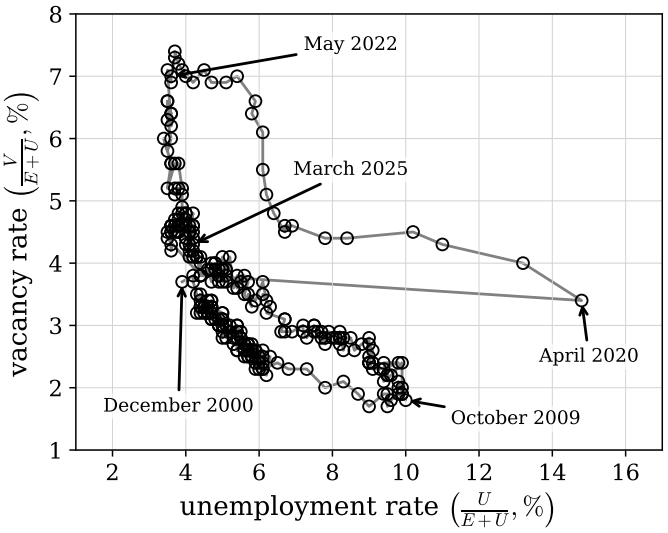


Figure 20.4: Beveridge curve in the United States.

Sources: JOLTS and CPS.

20.4.2 Market equilibrium with an endogenous vacancy creation

Let us consider how the vacancy level v_t is determined. Here, we assume the production is conducted by a match between one firm (vacancy) and one worker. We abstract from capital stock for the time being and assume that the match between a firm and a worker can produce z_t units of goods. We further assume that, as in the standard real business cycle (RBC) model, there is an aggregate shock to productivity. Therefore, z_t can vary stochastically over time. As in the RBC models, we assume that z_t follows a first-order Markov process.

It turns out that we can formulate the model recursively and the relevant state variable in the general equilibrium is only z. Assume that the firm discounts the future profit with the discount factor $\beta \in (0, 1)$. The Bellman equation for a firm that has already matched with a worker is

$$J(z) = z - w(z) + \beta \mathbb{E}[(1 - \sigma)J(z') + \sigma V(z')], \tag{20.7}$$

where $J(z)$ is the value of a matched firm. The flow value $z - w(z)$ is profit, where $w(z)$ is the wage paid to the worker when the aggregate productivity is z. The parameter $\beta \in (0, 1)$ is the discount factor of the firm (which will be identical to the worker's discount factor), and $\mathbb{E}[\cdot]$ indicates the expected value (conditional on the current period information). The prime ($'$) indicates the next period variable. $V(z)$ represents the value of a vacancy.

The Bellman equation for a vacant firm is

$$V(z) = -\kappa + \beta \mathbb{E}[\lambda_f(\theta)J(z') + (1 - \lambda_f(\theta))V(z')], \tag{20.8}$$

where $\kappa > 0$ is the cost of posting a vacancy. We also assume that anyone can set up a vacancy and enter the market ("free entry"). Thus, in equilibrium, the value of vacancy is driven down to zero:

$$V(z) = 0. \tag{20.9}$$

(20.8) and (20.9) imply

$$\frac{\kappa}{\lambda_f(\theta)} = \beta \mathbb{E}[J(z')]. \tag{20.10}$$

Intuitively, the cost of vacancy κ has to be equal to the expected value of the future filled job $\beta \mathbb{E}[J(z')]$ times the probability of finding a worker $\lambda_f(\theta)$.

To determine the equilibrium wage, we first have to consider the worker side. We assume that a worker is infinitely-lived, consumes what she receives every period, and has linear utility function with discount factor β (i.e., the same discount factor as the firm's):[4]

$$\mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t c_t \right].$$

The Bellman equation for an employed worker is

$$W(z) = w(z) + \beta \mathbb{E}[(1-\sigma)W(z') + \sigma U(z')], \tag{20.11}$$

where $W(z)$ is the value of an employed worker and $U(z)$ is the value of an unemployed worker. We assume that an unemployed worker receives a constant amount of goods $b < z_t$ (which has to hold for all possible values of z_t). b can be interpreted as home production or unemployment insurance payment. The Bellman equation for an unemployed worker is

$$U(z) = b + \beta \mathbb{E}[\lambda_w(\theta)W(z') + (1-\lambda_w(\theta))U(z')]. \tag{20.12}$$

Once a firm and a worker match, they are in a *bilateral monopoly* situation: the only possible seller (of the labor service) for the firm is the worker it matched with, and the only possible buyer for the worker is the firm she matched with. In such a situation, we cannot use the marginal principle to determine the wage because there is no competition. The match generates a surplus. In the current period, the match jointly generates z—if they separate, the worker can create b, and the firm can end up creating nothing. Therefore, it is jointly beneficial for the firm and the worker to be together because of the assumption $z > b$. Unless they are separated, this flow surplus $z - b$ is generated in the future as well. We assume that the firm and the worker split the surplus following the *Generalized Nash Bargaining* rule. The Nash Bargaining rule splits the surplus so that the Nash Product, which is the product of surpluses of each party (in our case, the firm and the worker), is maximized. The Generalized Nash Bargaining rule uses the weighted product instead, where the "weight" is represented as the exponent to each of the surpluses.

In our formulation, the Generalized Nash Bargaining solution solves

$$\max_w (\tilde{W}(w,z) - U(z))^{\gamma} (\tilde{J}(w,z) - V(z))^{1-\gamma},$$

[4]Because the firms in this economy can earn a profit (in particular, the aggregate profit is positive in the steady-state of the economy), there is a question of who receives the profit (i.e., the ownership of the firms). Here, we implicitly assume that the firms are owned by someone outside the economy who has the same discount factor as the worker. Alternatively, we can assume that the firm is owned by workers. As we will see later in this section, because only the *difference* of income between the employed worker and the unemployed worker matters for the equilibrium dynamics of unemployment, the same results go through if we assume that the workers' initial ownership of the firms (i.e., the stock holdings) is equal across workers. This result follows because, with linear utility, there are no reasons for the workers to trade stocks. In Section 20.8, we assume a closed economy and make the stock holding explicit.

where $\tilde{W}(w,z)$ is the value of an employed worker when the current wage is w. Note that the Bellman equation (20.11) assumes that the wage is the equilibrium value under z, $w(z)$. Here, we are allowing w to be any value. Therefore $\tilde{W}(w(z),z) = W(z)$ holds. Similarly, $\tilde{J}(w,z)$ is the value of a job matched with a worker when the wage is w. Here, the worker's surplus is $\tilde{W}(w,z) - U(z)$, and the firm's surplus is $\tilde{J}(w,z) - V(z)$. The exponent $\gamma \in (0,1)$ represents the "weight" of the worker's surplus. It is often referred to as the "bargaining power" of the worker. By taking the first-order condition, one can show that w solves

$$(1-\gamma)(\tilde{W}(w,z) - U(z)) = \gamma(\tilde{J}(w,z) - V(z)). \tag{20.13}$$

The detailed derivation of (20.13) is in Appendix 20.A.1.

The six equations (20.7), (20.8), (20.9), (20.11), (20.12), and (20.13) define the equilibrium. These can be rearranged to obtain a difference equation on θ_t:[5]

$$\frac{\kappa}{(1-\gamma)\lambda_f(\theta_t)} = \beta \mathbb{E}\left[z_{t+1} - b + \frac{1-\sigma-\gamma\lambda_w(\theta_{t+1})}{1-\gamma} \frac{\kappa}{\lambda_f(\theta_{t+1})} \right]. \tag{20.14}$$

When z is constant over time, the steady-state value of θ_t (denote it $\bar{\theta}$) solves

$$\frac{\kappa}{(1-\gamma)\lambda_f(\bar{\theta})} = \beta \left[z - b + \frac{1-\sigma-\gamma\lambda_w(\bar{\theta})}{1-\gamma} \frac{\kappa}{\lambda_f(\bar{\theta})} \right]. \tag{20.15}$$

This equation (note that the right-hand side is decreasing in $\bar{\theta}$) determines $\bar{\theta} = \bar{v}/\bar{u}$, where $\bar{v}$ is the steady-state value of vacancy. Often this condition is called the job creation condition. The job creation condition, together with the Beveridge curve relationship (20.6) determine $\bar{v}$ and $\bar{u}$.

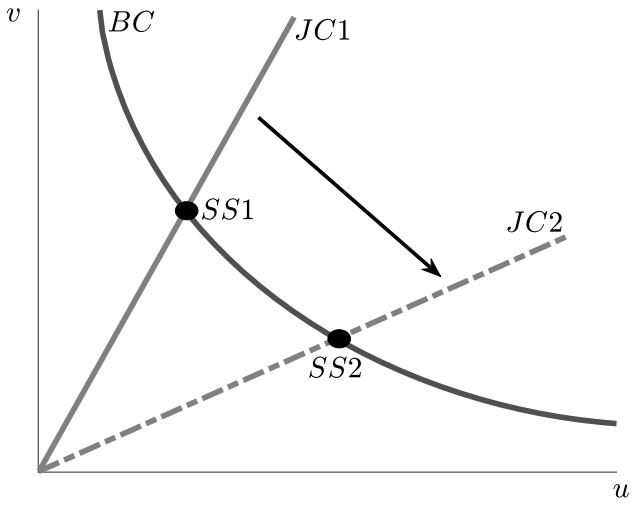


Figure 20.5: The determination of the steady state

[5] We can analyze the implications on equilibrium wages using the same set of equations. The analysis of the wages is in Appendix 20.A.2.

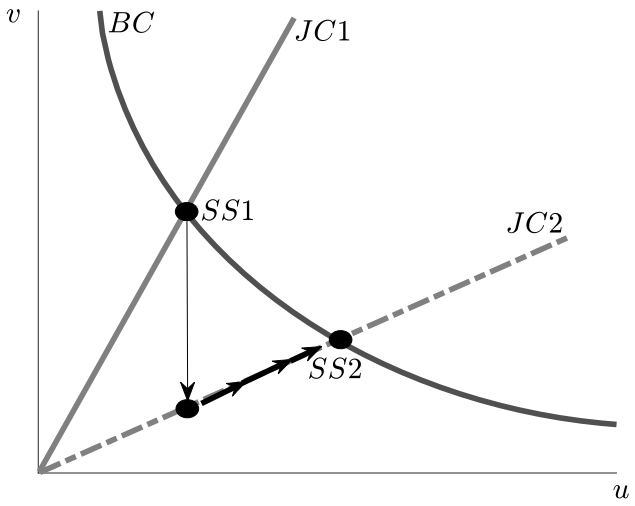


Figure 20.6: The transition dynamics of the DMP model.

Figure 20.5 describes the determination of v and u in the steady state. The BC curve represents (20.6), which describes the steady-state relationship between u and v. The straight lines $JC1$ and $JC2$ represent the relationship between u and v that correspond to different values of θ. When $JC1$ represents the value of θ that satisfy (20.15), the steady-state values of u and v are at $SS1$. When z goes down, from (20.15) we can see $\bar{\theta}$ also goes down. The JC line shifts from $JC1$ to $JC2$. In the new steady-state ($SS2$), v is smaller and u is larger.

The transition dynamics are also easy to analyze. Consider an unanticipated one-time permanent decline in z. Because the decline is permanent, the new job creation equation holds with the new steady-state value of θ. In other words, the equation (20.15) holds with the new value of $\bar{\theta}$. In Figure 20.6, the new value of $\bar{\theta}$ is represented by the new line $JC2$. The jump to the new value of $\bar{\theta}$ is immediate. Using equation (20.14), it is possible to show that θ_t would diverge away from the new steady state (and eventually violate the economy's feasibility) unless θ_t immediately jumps to the new steady-state value. Because u cannot jump, v immediately drops so that v/u becomes the new value of $\bar{\theta}$. After that, the economy gradually converges to the new steady state ($SS2$) along the $JC2$ line.

20.4.3 Efficiency

Unemployment has been considered one of the most important macroeconomic problems. Various policies have been proposed and implemented to reduce unemployment. Before considering policies, however, it is critical to know what kind of inefficiencies exist in the economy and whether unemployment in the market equilibrium is too high or too low compared to the social optimum.

One can think of the DMP model as describing the firm's investment problem through vacancy posting. We will show that there are two inefficiencies in making the investment decision.[6] First, because the firm is the only entity that makes the investment, the reward

[6]The exposition below closely follows Fukui and Mukoyama (2024).

has to be captured solely by the firm. In the DMP model, this is not the case. The surplus from matching is divided by the firms and the workers by the Generalized Nash Bargaining. This fact implies that the firm's reward from investment is "taxed," and the number of vacancies is inefficiently low. Second, a firm's decision imposes externalities on other firms and workers. Here, because the only entities that are making decisions are firms, what matters is the externality imposed on other firms. The externality imposed on the other firms is the number of vacancies times the change in the matching probability for each of the other firms, that is,

$$v \times \frac{\partial}{\partial v}\left(\frac{M(u,v)}{v}\right) = M_2(u,v) - \frac{M(u,v)}{v}. \tag{20.16}$$

The first term on the right-hand side, $M_2(u,v)$, is the number of matches created by the marginal vacancy. The second term, $M(u,v)/v$, is the private perception of the likelihood of a new match for a vacancy-posting firm. The difference is the externality. In other words, the externality is the difference between the marginal increase in the number of matches and the average number of matches per vacancy. Because the externality is negative, this inefficiency leads to too many vacancies in the market equilibrium. The balance of the two inefficiencies determines the overall effect.

There are similar inefficiencies in valuing the outcome of investment. The social value of moving a worker from unemployment to employment can be different from the private value. The social value takes the negative externality that an unemployed worker imposes on other unemployed workers. Because of this externality, the market equilibrium overvalues the opportunity cost of employment. At the same time, in the market equilibrium, only the opportunity cost for the worker is taken into account, and thus the opportunity cost is undervalued. The eventual outcome depends on the balance of these two inefficiencies.

To start the formal analysis, let us formulate the social planner's problem when the social planner is subject to the same labor market frictions as in the market equilibrium. This type of problem is often called the problem that solves the "constrained efficient" solution. The social planner maximizes the social welfare

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(z_t e_t + b(1-e_t) - \kappa v_t)\right],$$

where $\mathbb{E}_0[\cdot]$ is the expected value at time 0. The first term is the production by matched worker-firm, the second is the unemployed workers' home production, and the third is the vacancy-posting cost. Using $v_t = \theta_t u_t = \theta_t(1-e_t)$, let us write the social planner's problem as

$$\max_{\theta_t, e_{t+1}} \ \mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(z_t e_t + b(1-e_t) - \kappa\theta_t(1-e_t))\right]$$

subject to

$$e_{t+1} = (1-\sigma)e_t + \lambda_w(\theta_t)(1-e_t).$$

Let μ_t be the Lagrange multiplier of the constraint. Then, the Lagrangian is

$$L = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t(z_t e_t + b(1-e_t) - \kappa\theta_t(1-e_t)) + \sum_{t=0}^{\infty}\mu_t((1-\sigma)e_t + \lambda_w(\theta_t)(1-e_t) - e_{t+1})\right].$$

The first-order condition on θ_t is

$$\beta^t \kappa(1-e_t) = \mu_t \lambda_w'(\theta_t)(1-e_t)$$

and therefore

$$\kappa = \lambda_w'(\theta_t)\beta\hat{\mu}_t \tag{20.17}$$

holds, where $\hat{\mu}_t$ is defined as

$$\hat{\mu}_t \equiv \frac{\mu_t}{\beta^{t+1}}. \tag{20.18}$$

Because μ_t is the time-0 social value of increasing e_{t+1} by one unit (taking expectations at time t), $\hat{\mu}_t$ is the concurrent value of one unit of employment at time $t+1$ (taking expectations at time t). The left-hand side of (20.17), κ, is the social cost of increasing θ_t by one unit. By increasing θ_t by one unit, employment increases by $\lambda_w'(\theta_t)$ units, so the right-hand side of (20.17) is the social benefit of increasing θ_t from the viewpoint of the time when the vacancy is created.

The first-order condition on e_{t+1} is

$$\mu_t = \mathbb{E}_t\left[\beta^{t+1}(z_{t+1} - b + \kappa\theta_{t+1}) + \mu_{t+1}(1-\sigma-\lambda_w(\theta_{t+1}))\right].$$

Here, $\mathbb{E}_t[\cdot]$ is the expectation taken at period t. It can be rewritten as

$$\hat{\mu}_t = \mathbb{E}_t\left[z_{t+1} - b + \kappa\theta_{t+1} + \beta\hat{\mu}_{t+1}(1-\sigma-\lambda_w(\theta_{t+1}))\right]. \tag{20.19}$$

The two equations (20.17) and (20.19) characterize the socially optimal outcome, represented by θ_t and $\hat{\mu}_t$ for $t = 0, 1, \ldots$. The term $\beta\hat{\mu}_{t+1}\lambda_w(\theta_{t+1})$ corresponds to the opportunity cost of employment: a worker has to forgo an opportunity of a possible new match. However, having one more unemployed worker imposes an externality on other workers. The term $\kappa\theta_{t+1}$ corrects for this externality.

Let us derive the market equilibrium equations that correspond to these two. First, define the *expected surplus of a match* as

$$S_t \equiv \mathbb{E}_t[W(z_{t+1}) - U(z_{t+1}) + J(z_{t+1}) - V(z_{t+1})],$$

where the expectation is taken at time t. From (20.10) and (20.13) (which implies $(1-\gamma)S_t = \mathbb{E}_t[J(z_{t+1})]$), we obtain

$$\kappa = \beta(1-\gamma)\lambda_f(\theta_t)S_t. \tag{20.20}$$

From (20.7), (20.11), and (20.12) all for time $t+1$ (and taking expectations at time t), and using (20.9) and (20.13) (which implies $\mathbb{E}_{t+1}[W(z_{t+2}) - U(z_{t+2})] = \gamma S_{t+1}$),

$$S_t = \mathbb{E}_t\left[z_{t+1} - b + \beta S_{t+1}(1-\sigma-\gamma\lambda_w(\theta_{t+1}))\right]. \tag{20.21}$$

In the final term on the right-hand side, $\beta S_{t+1}\lambda_w(\theta_{t+1})$, is multiplied by γ because only the opportunity cost of a match on the worker side is taken into account in the market equilibrium.

To make a comparison between the social planner's solution and the market equilibrium, first, define the elasticity of the firm's worker-finding probability as

$$\eta(\theta) \equiv -\frac{\theta\lambda_f'(\theta)}{\lambda_f(\theta)}.$$

From $\lambda_w(\theta) = \theta\lambda_f(\theta)$, one can derive

$$\lambda_w'(\theta) = \lambda_f(\theta)(1 - \eta(\theta)).$$

This relationship enables us to rewrite the social planner's first-order condition (20.17) as

$$\kappa = \beta(1 - \eta(\theta_t))\lambda_f(\theta_t)\hat{\mu}_t. \tag{20.22}$$

This equation can be used to rewrite (20.19) as

$$\hat{\mu}_t = \mathbb{E}_t\left[z_{t+1} - b + \beta\hat{\mu}_{t+1}(1 - \sigma - \eta(\theta_{t+1})\lambda_w(\theta_{t+1}))\right]. \tag{20.23}$$

Comparing (20.20) against (20.22) and (20.21) against (20.23), one can see that the social planner's solution $(\theta_t, \hat{\mu}_t)$ and the equilibrium outcome (θ_t, S_t) are equivalent if

$$\eta(\theta) = \gamma \tag{20.24}$$

for all θ. This condition is often called the *Hosios condition* (Hosios, 1990). For the firm's investment incentive, the market equilibrium "taxes" the vacancy-creation incentive to correct for the negative externality imposed on the other firms. The "tax rate" γ has to be larger when $\eta(\theta)$ is larger because, when $\eta(\theta)$ is large, the change in other firms' matching probability $\lambda_f(\theta)$ due to the firm's vacancy posting is larger. One can also see that

$$\eta(\theta)\lambda_f(\theta) = -\theta\lambda_f'(\theta) = \theta\frac{d}{d\theta}\left(\frac{M(1,\theta)}{\theta}\right) = M_2(u,v) - \frac{M(u,v)}{v},$$

which implies that the term $\eta(\theta)\lambda_f(\theta)$ is indeed the externality imposed to the other firms, derived in (20.16).[7] For the valuation of worker employment, the Hosios condition ensures that the externality that an unemployed worker imposes on other unemployed workers corresponds to the under-valuation of the opportunity cost of employment in the market equilibrium.

Suppose the Hosios condition (20.24) does not hold. For example, suppose that $\eta(\theta) > \gamma$ for all θ. It can be shown that, in this case, the value of θ in the market equilibrium is too large compared to the constrained-efficient outcome. In this situation, the equilibrium unemployment rate is too low compared to the social optimum. A policy that reduces the unemployment rate, therefore, can lower the social welfare in this situation. In contrast, if $\eta(\theta) < \gamma$, the equilibrium unemployment rate is too high. In this situation, there is room for improving social welfare through social policies targeting lower unemployment.

20.5 Labor market facts, once again

The modern macroeconomic study of the labor market considers the gross flows behind the movement of stocks, such as unemployment in Figure 20.1. In fact, the model we presented in Sections 20.3 and 20.4 provides the analysis of gross flows between the state E (employment) and the state U (unemployment).

Empirical studies typically focus on three states of the labor market, E (employment), U (unemployment), and N (not in the labor force). Thus we can consider six flows across these states, as we can see in Figure 20.7.

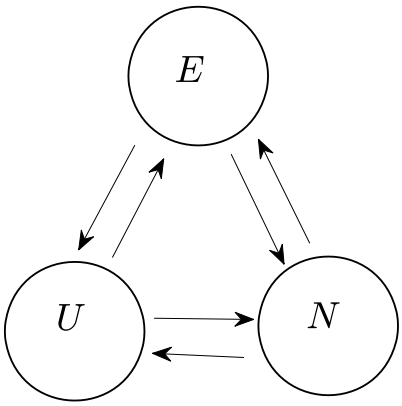


Figure 20.7: Flows among three states.

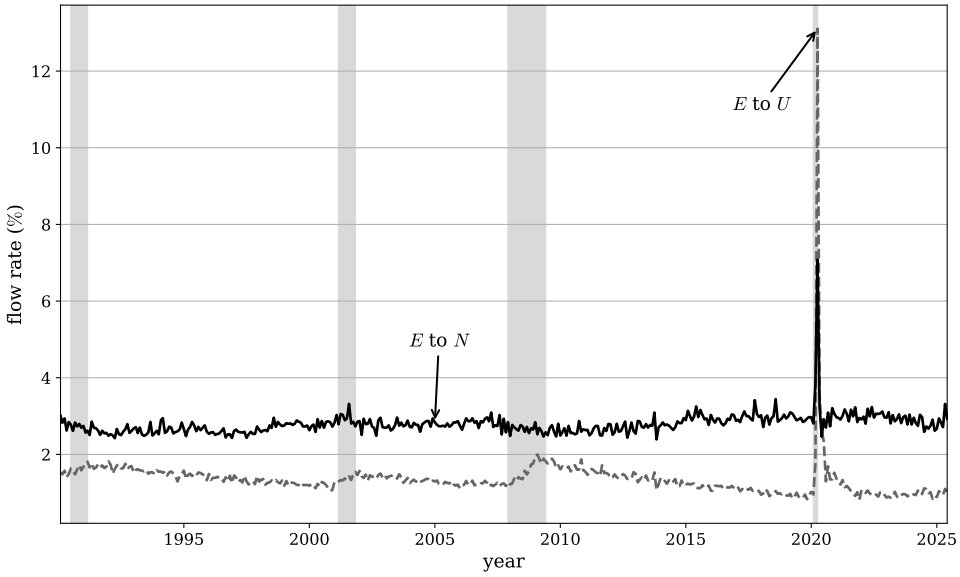


Figure 20.8: Flow rates out of E.

Sources: CPS (Fallick and Fleischman, 2004).

Figures 20.8, 20.9, and 20.10 plot these flow rates, from Fallick and Fleischman (2004).[8] For example, the E to U flow rate in Figure 20.8 plots the fraction of employed workers that flow into U in the following month. One can see how the gross flows influence the movement of stocks. For example, U increases in recessions because the inflows from E and N into U go up and the outflows to E and N go down.

These stylized facts provide important information for building the models of unemployment. Here, we point out two simple observations. First, the flow rate from U to E is

[7]The first equation uses the definition of $\eta(\theta)$, and the second equation uses the definition of $\lambda_f(\theta)$. For the third equation, the property $M_2(u,v) = M_2(1,\theta)$, which can be derived by differentiating both sides of $M(u,v) = uM(1,v/u)$ with respect to v, is used.

[8]The data is from `https://www.federalreserve.gov/pubs/feds/2004/200434/200434abs.html`. Seasonal adjustment is made using X-13 ARIMA-SEATS from the U.S. Census Bureau.

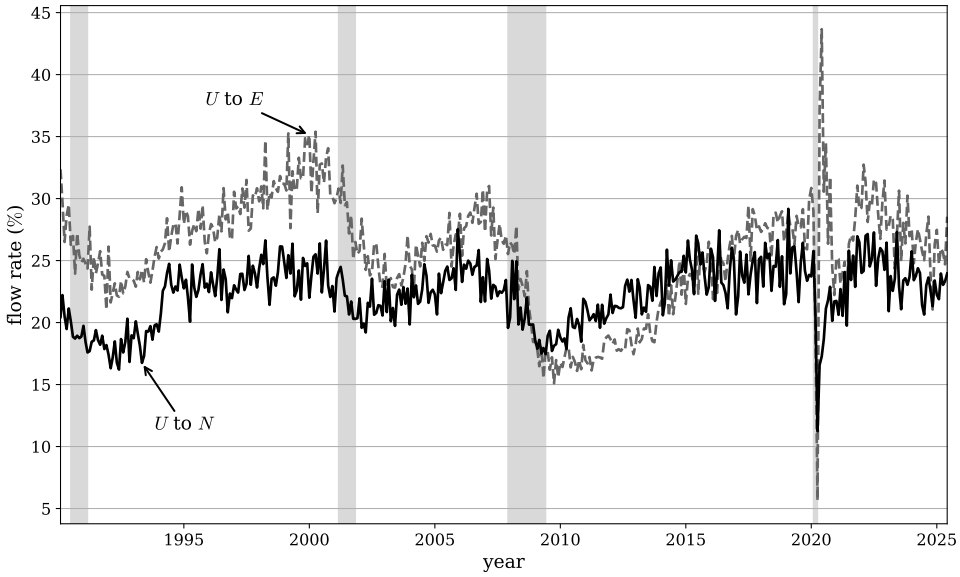


Figure 20.9: Flow rates out of U.

Sources: CPS (Fallick and Fleischman, 2004).

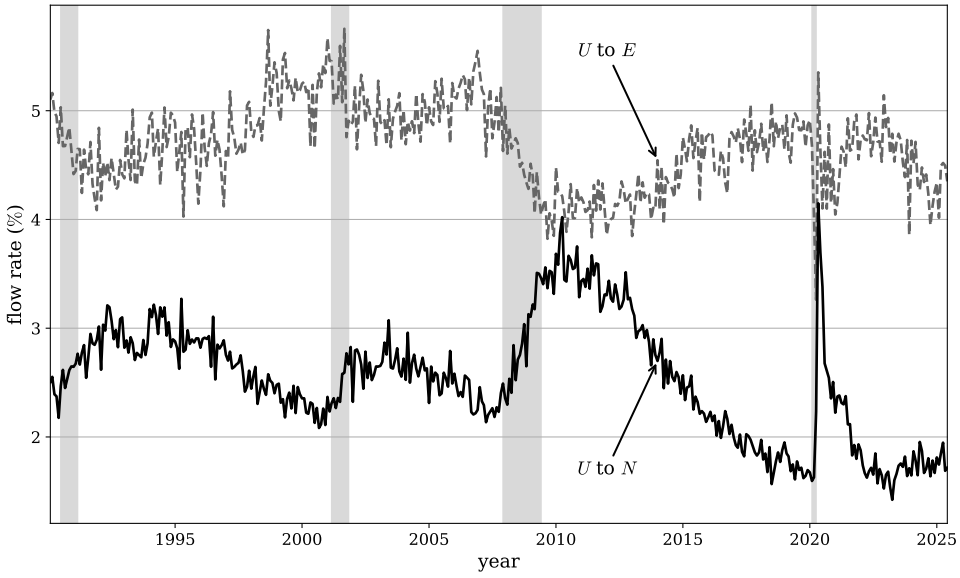


Figure 20.10: Flow rates out of N.

Sources: CPS (Fallick and Fleischman, 2004).

strongly procyclical. This property is consistent with the DMP model in Section 20.4. Second, the flow rate from E to U is strongly countercyclical. This fact implies the constant separation rate in the DMP model above is not consistent with the data. We will come back to this point in Section 20.7.

20.6 Unemployment volatility puzzle

We now examine the quantitative performance of the DMP model that we presented in Section 20.4. Let us go back to the difference equation (20.14). We will quantitatively evaluate the model (as in the analysis of the RBC models) by assigning functional forms and parameter values to the model.

Assume that the matching function is specified as the Cobb-Douglas form:

$$M(u,v) = \chi u^{\eta} v^{1-\eta}, \tag{20.25}$$

where $\eta \in (0,1)$. Then $\lambda_w(\theta) = \chi\theta^{1-\eta}$ and $\lambda_f(\theta) = \chi\theta^{-\eta}$. Note that we are using the same notation η as in Section 20.4.3. If we compute the (negative of) elasticity of $\lambda_f(\theta)$ with respect to θ, it is exactly the constant value η in this Cobb-Douglas form.

In addition, specify the process of z_t by

$$\hat{z}_{t+1} = \rho\hat{z}_t + \varepsilon_{t+1}, \tag{20.26}$$

where the hat ($\hat{}$) describes the log-deviation: $\hat{z}_t = \log(z_t) - \log(\bar{z})$ ($\bar{z}$ is the steady-state value of z). The parameter $\rho \in (0,1)$ represents the persistence of the shock, and the i.i.d. shock $\varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2)$. Here, $\log(\bar{z})$ is normalized to zero.

20.6.1 Log-linearized solution

We will solve this model by log-linearizing the solution around the steady state. Log-linearizing the equation (20.14) yields

$$\mathcal{A}\hat{\theta}_t = \mathbb{E}[\bar{z}\hat{z}_{t+1} + \mathcal{B}\hat{\theta}_{t+1}], \tag{20.27}$$

where

$$\mathcal{A} = \frac{\kappa\bar{\theta}^{\eta}\eta}{(1-\gamma)\beta\chi}$$

and

$$\mathcal{B} = \frac{1-\sigma}{1-\gamma}\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi} + \frac{\gamma\kappa\bar{\theta}}{1-\gamma}.$$

Appendix 20.A.3 briefly describes the basic method of log-linearization. Further details can be found, for example, in Uhlig (2001).

To solve the difference equation (20.27) using the method of undetermined coefficients, first guess $\hat{\theta}_t = \mathcal{C}\hat{z}_t$, where $\mathcal{C}$ is the undetermined coefficient. Plugging this guess into (20.27) and using $\mathbb{E}[z_{t+1}] = \rho z_t$, we obtain

$$\mathcal{C} = \frac{\rho}{\mathcal{A} - \rho\mathcal{B}}.$$

Rearranging, we obtain the relationship between $\hat{\theta}_t$ and $\hat{z}_t$ as

$$\hat{\theta}_t = (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\sigma)\right) + \kappa\gamma\bar{\theta}\right]^{-1}\hat{z}_t. \tag{20.28}$$

Analytically, this result provides important insights into the model performance against the data. In particular, (20.28) implies (for a given $\bar{\theta}$), a small κ, a small γ, a small η, a large χ, a large ρ, a large β, and a small σ makes the response of θ_t larger.

Table 20.1: Parameter values

Calibrated Parameters	Value
β	0.996
ρ	0.949
σ_ε	0.0065
σ	0.034
χ	0.45
b	0.4
γ	0.72
η	0.72

20.6.2 Calibration

Now we set the relevant parameter values. Let one period in the model be one month, so that we can use the U.S. labor market data and capture the fast labor market dynamics in the U.S. economy. Here, we assume the parameter values in Table 20.1.

The discount factor β is set at $0.947^{\frac{1}{12}} = 0.996$. The annual value of 0.947 is taken from the standard real business cycle literature (Cooley and Prescott, 1995).

The parameters for the process for $\hat{z}$, ρ, and σ_ε, are taken from Hagedorn and Manovskii (2008) and adjusted to a monthly frequency. Here, σ_ε is the standard deviation of ε_{t+1} in (20.26), assuming that ε_{t+1} follows a normal distribution.

The values $b = 0.4$, $\eta = 0.72$, and $\gamma = 0.72$ follows Shimer (2005). Furthermore, following Shimer (2005), the parameter value for κ is set so that the equation (20.15) is satisfied in the steady state with $\bar{\theta} = 1$. The values of χ and σ also follow Shimer (2005).[9]

20.6.3 Quantitative results

With the given parameter values and the equations (20.4) and (20.28), we can simulate the model by randomly generating the series of $\hat{z}_t$ following (20.26). Table 20.2 is the summary of the U.S. data that we will compare the model against. Here, z measures labor productivity. All are originally from monthly data but averaged to quarterly data, and logged and HP-filtered with the smoothing parameter of 1600. The table is taken from Hagedorn and Manovskii (2008).

Table 20.3 is the statistics from the model-generated data. The model generates the right correlations between variables, but the magnitude of the fluctuations in u, v, and θ is too small compared to the data. This discrepancy is often referred to as the *unemployment volatility puzzle* or the *labor market volatility puzzle* (or the "Shimer puzzle," after Shimer

[9]These values are larger than the UE flow rates and the EU flow rates in Figures 20.8 and 20.9. This discrepancy is because Shimer (2005) calculates the outflow from and the inflow into U (which include the flows in and out of N), instead of UE and EU flow rates. His methodology does not require matching workers from one month to the next; instead, it utilizes the duration distribution of unemployment.

Table 20.2: Summary statistics for quarterly U.S. data

		u	v	v/u	z
Standard Deviation		0.125	0.139	0.259	0.013
Quarterly Autocorrelation		0.870	0.904	0.896	0.765
Correlation Matrix	u	1	-0.919	-0.977	-0.732
	v	—	1	0.982	0.460
	v/u	—	—	1	0.967
	z	—	—	—	1

Table 20.3: Model statistics

		u	v	v/u	z
Standard Deviation		0.005	0.016	0.020	0.013
Quarterly Autocorrelation		0.826	0.700	0.764	0.765
Correlation Matrix	u	1	-0.839	-0.904	-0.804
	v	—	1	0.991	0.972
	v/u	—	—	1	0.961
	z	—	—	—	1

(2005)). The model's inability to match the magnitude of labor market fluctuations triggered an extensive body of research in early 2000s.

Intuitively, there are two reasons, corresponding to benefits and costs of hiring, for the quantitatively small response. First, the benefit of hiring a worker is procyclical, but the magnitude is not large. One reason is that the wage increases in booms, and it weakens the response of profit to the productivity shock. Second, the cost of hiring a worker, $\kappa/\lambda_f(\theta)$ moves together with θ. In booms, θ increases, and this increase in cost dampens the firm's response to a positive productivity shock.

20.6.4 Rigid wages

Many possible "solutions" are proposed for the unemployment volatility puzzle. Although there is no clear consensus among researchers in terms of which proposed "solution" is the most plausible one, here we highlight the role of rigid wages. As we explained above, the response of wages to productivity shocks dampens the volatility of profits. Rigid wages would make the benefit of creating a vacancy more volatile.

Empirically, there have been many studies about wage rigidity, both nominal and real.[10] Even without search frictions, rigid wages can generate unemployment by preventing the labor market from clearing. Here, wage rigidity changes the incentive for the firms to hire workers in booms and recessions.[11]

Instead of the Generalized Nash Bargaining, here we assume that wages are rigid at the

[10] See, for example, McLaughlin (1994) and Elsby and Solon (2019).

[11] The role of wage rigidity in this context was first explored by Hall (2005) and Shimer (2005).

steady-state value $w = \bar{w}$. Combining (20.7) with (20.10), we obtain

$$\frac{\kappa}{\lambda_f(\theta_t)} = \beta \mathbb{E}\left[z_{t+1} - \bar{w} + \frac{(1-\sigma)\kappa}{\lambda_f(\theta_{t+1})} \right].$$

Log-linearizing,

$$\hat{\theta}_t = \left[\frac{\kappa \bar{\theta}^{\eta} \eta}{\chi} \left(\frac{1}{\rho\beta} - (1-\sigma) \right) \right]^{-1} \hat{z}_t. \tag{20.29}$$

With the same calibration, the model outcome can be computed as in Table 20.4. Unemployment fluctuations are of the same magnitude as the data. Comparing (20.28) and (20.29) reveals two factors that the wage rigidity can make the firm's profit more volatile. First, the latter is not multiplied by $(1-\gamma)$, indicating that the additional gain in the surplus is no longer shared between the worker and the firm, and firm can receive all the additional gain. Second, the term $\kappa\gamma\bar{\theta}$, which represents the improvement of the worker's bargaining position due to the rise in the future job finding probability, is absent when wages are rigid.

Therefore, one can see that a (real) wage rigidity can address the volatility puzzle. We note, however, the sources and magnitude of the wage rigidity still remains an active area of research.

Table 20.4: Model statistics with fixed wages

		u	v	v/u	z
Standard Deviation		0.115	0.329	0.425	0.013
Quarterly Autocorrelation		0.825	0.693	0.763	0.765
Correlation Matrix	u	1	−0.791	−0.881	−0.784
	v	—	1	0.986	0.969
	v/u	—	—	1	0.961
	z	—	—	—	1

20.7 Endogenous separation

The previous sections focused on the role of the fluctuations in job-finding probability. As we saw in Section 20.5, both job finding and separation rates are both strongly cyclical.[12] In particular, at the onset of recessions, the increase in the separation rate tends to cause a sharp increase in the unemployment rate. In Figure 20.8, the *EU* flow rate increases in recessions, and the magnitude of the increase depends on the severity of the recession, suggesting that the separation rate changes endogenously with the business cycle. In this section, we extend the basic model in Section 20.4 to allow for endogenous separation.

[12] Fujita and Ramey (2009) is an earlier work that emphasize the importance of the separation margin in unemployment fluctuations.

20.7.1 Formulation

Instead of facing an exogenous separation shock, the firm has to pay a cost for maintaining the match, $c(\sigma)$. Now σ is a choice variable for the firm, but the cost increases if the firm wants to make the separation probability small, that is, $c'(\sigma) < 0$.

The matched firm's Bellman equation is now

$$J(z) = \max_{\sigma} \; z - w(z) - c(\sigma) + \beta \mathbb{E}\left[(1-\sigma)J(z') + \sigma V(z')\right].$$

The rest of the equilibrium conditions ((20.8), (20.9), (20.11), (20.12), and (20.13)) are the same as in the Section 20.4. The optimal value of σ is now a function of z (denote it as $\sigma(z)$). The unemployment dynamics is, therefore,

$$u_{t+1} = (1 - \lambda_w(\theta_t(z_t)))u_t + \sigma(z_t)(1-u_t).$$

The first-order condition for σ is, using (20.9),

$$-c'(\sigma) = \beta \mathbb{E}[J(z')].$$

From (20.10), this equation implies

$$-c'(\sigma) = \frac{\kappa}{\lambda_f(\theta)}. \tag{20.30}$$

The job creation condition can be derived in the same manner as in the exogenous separation case:

$$\frac{\kappa}{(1-\gamma)\lambda_f(\theta_t)} = \beta \mathbb{E}\left[z_{t+1} - b - c(\sigma(z_{t+1})) + \frac{1 - \sigma(z_{t+1}) - \gamma\lambda_w(\theta_{t+1})}{1-\gamma}\frac{\kappa}{\lambda_f(\theta_{t+1})}\right]. \tag{20.31}$$

The equation (20.31) determines the dynamics of θ_t. Here, $\sigma(z)$ is determined in equilibrium by (20.30) (because θ is a function of z, σ is also a function of z).

20.7.2 Log-linearized system

Once again, we work with a log-linearized system. First, let the maintenance cost function be

$$c(\sigma) = \phi \sigma^{-\xi},$$

where $\phi > 0$ and $\xi > 0$ are parameters. Assume that the matching function takes the form (20.25). The derivation of the log-linearized system is similar to the exogenous separation case and detailed in Appendix 20.A.4.

First, guess the log-linearized relationship between θ_t and z_t as

$$\hat{\theta}_t = \mathcal{G}\hat{z}_t. \tag{20.32}$$

Then (20.30) can be log-linearized to

$$\hat{\sigma}(z_t) = -\frac{\eta}{\xi+1}\mathcal{G}\hat{z}_t. \tag{20.33}$$

Thus the log-deviation of the cost is (using the shortened notation of $c(z) = c(\sigma(z))$)

$$\hat{c}(z_t) = \frac{\xi\eta}{\xi+1}\mathcal{G}\hat{z}_t. \tag{20.34}$$

Using (20.33) and (20.34), (20.31) can be The coefficient $\mathcal{G}$ can be solved as

$$\mathcal{G} = \frac{\Theta}{\Gamma},$$

where

$$\Theta \equiv (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{z}$$

and

$$\Gamma \equiv 1 + (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{c}\frac{\xi\eta}{\xi+1}\left(1 - \frac{1}{1-\gamma}\right).$$

where $\bar{\sigma}$ is the steady-state value of σ and $\bar{c} = \phi\bar{\sigma}^{-\xi}$ is the steady-state value of the maintenance cost.

20.7.3 Calibration and quantitative results

Parameters β, ρ, σ_ε, χ, b, γ, and η are set at the same value as in the previous section. For σ, we set the other parameters so that $\bar{\sigma} = 0.034$ matches the average separation rate in the U.S. data.

The newly-introduced specification is the maintenance cost function, $c(\sigma) = \phi\sigma^{-\xi}$. (20.32), (20.33), and $\lambda_w(\theta) = \chi\theta^{1-\eta}$ implies the ratio of the standard deviations

$$\frac{\mathrm{std}(\hat{\lambda}_w)}{\mathrm{std}(\hat{\sigma})} = \frac{(1-\eta)(1+\xi)}{\eta}$$

Krusell et al. (2017, Table 8) indicates the ratio of the standard deviations for the EU flow rate and the UE flow rate is close to 1. Thus, for $\eta = 0.72$, we set $\xi = 1.6$.

As in Section 20.6.2, we set the steady-state value of θ as 1, and for a given κ, we can determine the steady-state value of ϕ from $\bar{\sigma} = 0.034$. Thus $c(\bar{\sigma})$ can be expressed in κ, and as in Section 20.6.2, we can set the value of κ from the steady-state version of (20.31).

The result is in Table 20.5. The fluctuation of u is still quantitatively very small compared to the data, although it is larger than the constant σ case, thanks to the movement in σ.

20.7.4 Rigid wages

Once again, we examine the situation where wages are rigid. Following the same steps as those in the Generalized Nash Bargaining case, it can be shown that

$$\frac{\kappa}{\lambda_f(\theta_t)} = \beta\mathbb{E}\left[z_{t+1} - \bar{w} - c(z_{t+1}) + \frac{(1-\sigma(z_{t+1}))\kappa}{\lambda_f(\theta_{t+1})}\right]$$

Table 20.5: Model statistics with endogenous σ

		u	v	v/u	z
Standard Deviation		0.010	0.011	0.021	0.013
Quarterly Autocorrelation		0.862	0.623	0.764	0.765
Correlation Matrix	u	1	-0.893	-0.969	-0.909
	v	—	1	0.976	0.957
	v/u	—	—	1	0.961
	z	—	—	—	1

characterizes the dynamics of θ_t. Log-linearizing this equation, we obtain

$$\hat{\theta}_t = \left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right)\right]^{-1}\hat{z}_t.$$

It turns out that the obtained outcome is identical to (20.29). The terms with endogenous $c(z_t)$ and $\sigma(z_t)$, which are new elements here, exactly cancel out. The log-linearized equations for the dynamics of $\sigma(z_t)$ and $c(z_t)$, (20.33) and (20.34), are the same as in the Generalized Nash Bargaining case.

The results are in Table 20.6. The model can replicate the large fluctuations in unemployment. In fact, the fluctuations in u are larger than in the data, suggesting that even a less extreme form of wage rigidity can generate substantial fluctuations in u in this case.

Table 20.6: Model statistics with endogenous σ and fixed wages

		u	v	v/u	z
Standard Deviation		0.217	0.232	0.433	0.013
Quarterly Autocorrelation		0.852	0.609	0.763	0.765
Correlation Matrix	u	1	-0.854	-0.960	-0.901
	v	—	1	0.966	0.948
	v/u	—	—	1	0.961
	z	—	—	—	1

20.8 Labor market frictions and the neoclassical growth model

As the final section of this chapter, we connect the DMP model to the main workhorse model of this book: the neoclassical growth model.[13] The important changes are (i) concave utility with an explicit consumption-saving problem and (ii) the use of capital (in addition to labor) in production. In addition, as discussed in the footnote 4 briefly, the earlier sections

[13]The model in this section follows Krusell, Mukoyama, and Şahin (2010, Appendix O). The earlier papers incorporating the search and matching framework into the neoclassical growth model include Merz (1995) and Andolfatto (1996).

implicitly assume that all firms are owned by someone outside the economy. In this section, we consider a closed economy, and therefore the profit income from the firm ownership is made explicit.

20.8.1 The baseline model with Generalized Nash Bargaining

Imagine that there are consumers on the unit square. The mass of consumers is one. The consumers are indexed by (i, j), where $i \in [0, 1]$ and $j \in [0, 1]$. The index i indicates the family the consumer belongs to. The family i, therefore, has members (indexed by j) on a unit interval. Families are identical to each other, and therefore, we will consider the *representative family*. Within each family, the members insure each other. That is, although some members are employed and others are unemployed, the income is pooled at the family level, and each member consumes the same amount (here, we do not consider disutility from work). Thus we have families that are homogeneous, and the family members are identical within each family. Below we will consider the decision of the representative family. Because the consumption of each family member is identical, the "family head" only needs to think about the representative member of the family.

Assume that each family member's utility is (because we consider the representative member of the representative family, we omit the indices i and j below)

$$\mathbb{E}_0 \left[\sum_{t=0}^{\infty} \mathbf{U}(c_t) \right],$$

where $\mathbb{E}_0[\cdot]$ is the expectation taken at time 0, c_t is and $\mathbf{U}(\cdot)$ is an increasing and concave period utility function.

The budget constraint for the family is

$$c_t + k_{t+1} = (1 + r_t - \delta)k_t + (1 - u_t)w_t + u_t b + d_t,$$

where k_t is the capital stock holding, r_t is the rental rate of capital, $\delta \in (0, 1]$ is the depreciation rate of capital stock, u_t is the unemployment rate, w_t is the wage per worker, b is the home production of unemployed workers, and d_t is the dividend from the firm.

The labor market setting is the same as in Section 20.4. In this section, we assume that production uses both capital and labor. Normalizing the labor input per match as 1, the output per match is assumed to be $z_t k_t^{\alpha}$, where $\alpha \in (0, 1)$. We assume that the capital is rented by firms from the families every period. Thus, the maximization problem for the choice of capital input by the firm is

$$\max_{k_{f,t}} \; z_t (k_{f,t})^{\alpha} - r_t k_{f,t}.$$

The optimal capital input satisfies

$$\alpha z_t (k_{f,t})^{\alpha - 1} = r_t.$$

In equilibrium, there is a mass $(1 - u_t)$ of matches, which equally divides the total capital stock in the economy. Therefore, r_t in the equilibrium is

$$r(z_t, K_t, u_t) = \alpha z_t \left(\frac{K_t}{1 - u_t} \right)^{\alpha - 1}.$$

Below, let us call $X_t \equiv (z_t, K_t, u_t)$ for the shorthand. The surplus per match (denoted by z_t in Section 20.4) is now

$$y(X_t) \equiv z_t \left(\frac{K_t}{1-u_t}\right)^{\alpha} - r(X_t)\left(\frac{K_t}{1-u_t}\right) = (1-\alpha) z_t \left(\frac{K_t}{1-u_t}\right)^{\alpha}.$$

Firms also solve the dynamic problem of vacancy posting, as in the standard DMP model. Because the representative family's utility function is not linear, the discount factor can be different from β.

For the purpose of exposition, we divide the equilibrium of this model into two blocks: the consumption-saving block and the labor market block. The consumption-saving problem of the representative family can be written as the Bellman equation

$$\mathbf{V}(k, X) = \max_{c, k'} \ \mathbf{U}(c) + \beta \mathbb{E}[\mathbf{V}(k', X')|z] \tag{20.35}$$

subject to

$$\begin{aligned} c + k' &= (1 + r(X) - \delta)k + (1-u)w(X) + ub + d(X), \\ K' &= \Omega(X), \end{aligned} \tag{20.36}$$

and

$$u' = (1 - \lambda_w(\theta(X)) + \sigma(1-u), \tag{20.37}$$

where prime ($'$) represents the next period variable. The family takes the rental rate $r(X)$, the wage $w(X)$, the dividend $d(X)$, the law of motion for aggregate capital $\Omega(X)$, and the labor-market tightness $\theta(X)$ as given (as functions of X). Later on, we will confirm that $w(X)$, $d(X)$, $\Omega(X)$, and $\theta(X)$ are functions of X. From the solution to this Bellman equation, we obtain the decision rules $c(k, X)$ and $k'(k, X)$. Because the families are identical, the equilibrium aggregate consumption is $C(X) = c(K, X)$ and the next period aggregate capital is $\Omega(X) = k'(K, X)$. Note that (20.37) implies we can express u' as a function of X, $u'(X)$.

From this information, we can express the state price (i.e., the price of an Arrow security) of the next period state z' when the current state is X as

$$Q(z', X) = \beta f(z'|z) \frac{\mathbf{U}'(C(z', \Omega(X), u'(X)))}{\mathbf{U}'(C(X))}. \tag{20.38}$$

Here, $f(z'|z)$ is the probability density of state z' given the current state z and $u'(X)$ represents the right-hand side of (20.37). The derivation is in Appendix 20.A.5. In summary, once we know the functions $w(X)$, $d(X)$, and $\theta(X)$, we can obtain the state price $Q(z', X)$, in addition to other functions that include $\Omega(X)$. Below, we show that once we have $Q(z', X)$ and $\Omega(X)$, we can obtain $w(X)$, $d(X)$, and $\theta(X)$ in the "labor market block" below. Then these five functions $(Q(z', X), \Omega(X), w(X), d(X), \theta(X))$ can be computed as a fixed point.

Thus suppose we know $Q(z', X)$ and consider the labor market. It works very similarly to the basic DMP model. A firm with a worker has a value $J(X)$, where

$$J(X) = y(X) - w(X) + \int Q(z', X)[(1-\sigma)J(X') + \sigma V(X')]dz'. \tag{20.39}$$

Here, we discount the future value with $Q(z', X)$ because it represents the price of the next period good (when the state is z') in terms of the current good. The derivation of (20.39) (and the other asset value equations) can be found in Appendix 20.A.6. The value of vacancy is

$$V(X) = -\kappa + \int Q(z', X)[\lambda_f(\theta(X))J(X') + (1 - \lambda_f(\theta(X))V(X')]dz'.$$

Here, the transition equation (20.36) and (20.37) are given, and the functions $w(X)$ and $\theta(X)$ are a part of the unknowns in this block. The free-entry condition, $V(X) = 0$, implies

$$\frac{\kappa}{\lambda_f(\theta(X))} = \int Q(z', X)J(X')dz'. \tag{20.40}$$

This equation is analogous to (20.10) in the basic DMP model.

On the worker side, from the family's viewpoint, a worker brings in a stream of income with a stochastically changing employment state. Thus, we can compute the value of having a worker with specific status for a family using the standard asset pricing theory (the "Lucas tree" model). The value of an employed worker is

$$W(X) = w(X) + \int Q(z', X)[(1 - \sigma)W(X') + \sigma U(X')]dz' \tag{20.41}$$

and the value of an unemployed worker is

$$U(X) = b + \int Q(z', X)[\lambda_w(\theta(X))W(X') + (1 - \lambda_w(\theta(X)))U(X')]dz'. \tag{20.42}$$

Because $J(X) - V(X)$ and $W(X) - U(X)$ are both linear in w, with the same procedure as in Section 20.4, the Generalized Nash Bargaining implies

$$(1 - \gamma)(W(X) - U(X)) = \gamma(J(X) - V(X)). \tag{20.43}$$

The Generalized Nash Bargaining here implies that the wage is indeed a function of X.

From (20.39), (20.41), (20.42), and $V(X) = 0$,

$$W(X)-U(X)+J(X) = y(X)-b+\int Q(z', X)[(1-\sigma-\lambda_w(\theta(X)))(W(X')-U(X'))+(1-\sigma)J(X')]dz'$$

Using (20.43),

$$\frac{J(X)}{1-\gamma} = y(X) - b + \int Q(z', X)J(X')\frac{1 - \sigma - \gamma\lambda_w(\theta(X))}{1-\gamma}dz'. \tag{20.44}$$

Moving one period forward, multiplying $Q(z', X)$ on both sides, integrating over z', and using (20.40) yields the job creation condition:

$$\frac{\kappa}{(1-\gamma)\lambda_f(\theta(X))} = \int Q(z', X)\left[y(X') - b + \frac{1 - \sigma - \gamma\lambda_w(\theta(X'))}{1-\gamma}\frac{\kappa}{\lambda_f(\theta(X'))}\right]dz'. \tag{20.45}$$

This condition confirms that the equilibrium θ can indeed be written as a function of X. From (20.39), (20.40), and (20.44), the wage can be solved as

$$w(X) = \gamma(y(X) - b) + b + \gamma\theta(X)\kappa. \tag{20.46}$$

The dividend is all firms' profit minus the vacancy cost. The number of filled jobs is $(1-u)$, and each job creates $y(X) - w(X)$ units of profit. The vacancy cost is $\kappa v = \kappa\theta(X)u$ because $\theta(X) = v/u$. Thus

$$d(X) = (1-u)(y(X) - w(X)) - \kappa\theta(X)u, \tag{20.47}$$

and once again, we confirm that $d(X)$ is a function of X.

As with the standard RBC models, there are several alternative methods to compute the equilibrium. The first is, as in the previous sections, to log-linearize the equilibrium conditions and solve for the coefficients.

The second method is to treat the equilibrium conditions as functional equations. For example, one method that can be employed is to first make a guess on $Q(z', X)$, then use the (20.45) to find the function $\theta(X)$ (one can use an iterative method—start from $\theta(X)$ on the right-hand side to obtain $\theta(X)$ in the left-hand side, etc.). Then we can compute $w(X)$ and $d(X)$ from (20.46) and (20.47). Using this information, the representative family's problem (20.35) can be solved using the standard techniques to solve the neoclassical growth models. Finally, $Q(z', X)$ is updated with (20.38). We iterate this process until convergence. The following simulation follows this latter method of computation.

The details of calibration and computation are in Appendix 20.A.7. Calibration is similar to Section 20.6.2. The only difference from Table 20.1 is that, in this model, the steady-state value of $y(X)$ is not 1. Thus we adjust the value of b so that it is 0.4 times the steady-state value of $y(X)$. As in Section 20.6.2, κ is endogenously calibrated. The utility is assumed to be a log function $\mathbf{U}(c) = \log(c)$. The production function parameter $\alpha = 0.4$ as in the standard RBC model (Cooley and Prescott, 1995). The value of δ in Cooley and Prescott (1995) is 0.012 in quarterly frequency, and thus we set $\delta = 0.004$.

Table 20.7: Model statistics with Generalized Nash Bargaining: labor market

		u	v	v/u	z
Standard Deviation		0.005	0.017	0.022	0.015
Quarterly Autocorrelation		0.819	0.688	0.755	0.763
Correlation Matrix	u	1	−0.831	−0.899	0.089
	v	—	1	0.991	−0.071
	v/u	—	—	1	−0.078
	z	—	—	—	1

Table 20.7 computes the labor market statistics as in the earlier sections. The results are overall in line with Table 20.3. The only noticeable difference is that $\text{corr}(z, u)$, $\text{corr}(z, v)$, and $\text{corr}(z, v/u)$ are close to zero (and the signs are different). The reason is that, in this section's model, the production per worker $y(X)$ is affected not only by z, but also by k and u. In fact, the correlations of u, v, and v/u with the labor productivity $y(X)$ are similar to these with z in Table 20.3.

Table 20.8: Model statistics with Generalized Nash Bargaining: business cycles

	Y	C	I	L	Y/L
Standard Deviation	0.014	0.003	0.059	0.0004	0.014
Correlation with Y	1	0.875	0.991	0.902	0.99992

Table 20.8 computes the standard business cycle statistics that are typically computed in the Real Business Cycle (RBC) literature. Similar to the labor market statistics, all variables are aggregated to quarterly frequency, logged, and HP-detrended (with the smoothing parameter $\lambda = 1,600$). The business cycle properties are overall similar to the standard RBC model: all C, I, L, and Y/L (here, L is computed as $1-u$) are strongly procyclical, I is more volatile than Y, and C is less volatile than Y. The only significant difference is that L fluctuates much less than Y. In this model, this outcome reflects the unemployment volatility puzzle in Section 20.6.

20.8.2 Rigid wages

Now consider the case with rigid wages. The consumer's problem is the same as the Generalized Nash Bargaining case, except that the wage is rigid. Different from Section 20.6.4, the output per worker $y(X)$ moves not only with z, but also with k and u. Because the movement of $y(X)$ is relatively large, it turns out that the flow profit for the firm sometimes becomes negative. To maintain a positive profit, this time we assume the wage to be

$$\tilde{w}(X) = \max\{\bar{w}, y(X)\}.$$

In our simulation, in the majority of the periods, the wage remains $\bar{w}$.

The Bellman equation is

$$\mathbf{V}(k, X) = \max_{c,k'} \ \mathbf{U}(c) + \beta \mathbb{E}[\mathbf{V}(k', X')|z] \tag{20.48}$$

subject to

$$c + k' = (1 + r(X) - \delta)k + (1-u)\tilde{w}(X) + ub + d(X),$$
$$K' = \Omega(X),$$
$$u' = (1 - \lambda_w(\theta(X)) + \sigma(1-u),$$

The state price $Q(z', X)$ can, again, be computed as (20.38). The Bellman equation for the matched job is

$$J(X) = y(X) - \tilde{w}(X) + \int Q(z', X)[(1-\sigma)J(X') + \sigma V(X')]dz'. \tag{20.49}$$

The free-entry condition remains the same as (20.40), and thus (20.49) can be rewritten as

$$\frac{\kappa}{\lambda_f(\theta(X))} = \int Q(z', X) \left[y(X') - \tilde{w}(X) + \frac{(1-\sigma)\kappa}{\lambda_f(\theta(X'))} \right] dz'. \tag{20.50}$$

Similar to (20.47), the dividend is

$$d(X) = (1-u)(y(X) - \tilde{w}(X)) - \kappa\theta(X)u, \tag{20.51}$$

The computation is similar to the Generalized Nash Bargaining case, except that now $\tilde{w}(X)$ does not move as much. First make a guess on $Q(z', X)$. Second, we can solve for $\theta(X)$ from (20.50). Third, $d(X)$ can be computed from (20.51). Using these information, we can solve (20.48) and update $Q(z', X)$. These steps are repeated until $Q(z', X)$ converges.

Table 20.9: Model statistics with rigid wages: labor market

		u	v	v/u	z
Standard Deviation		0.083	0.269	0.339	0.015
Quarterly Autocorrelation		0.818	0.671	0.744	0.763
Correlation Matrix	u	1	−0.811	−0.886	0.094
	v	—	1	0.990	−0.076
	v/u	—	—	1	−0.083
	z	—	—	—	1

The model calibration is the same as in Section 20.8.1. Table 20.9 describes the labor market statistics for the rigid wage case. As in Section 20.6.4, the response of v (and therefore u) to the productivity shock is magnified by the rigid wage. Similarly to Table 20.7, the correlations of u, v, and v/u with z are weak. Once again, this result comes from the fact that $y(X)$ also moves with k and u. When we compute the correlation of these variables with $y(X)$, the pattern of correlations is similar to the results in the basic model.

Table 20.10: Model statistics with rigid wages: business cycles

	Y	C	I	L	Y/L
Standard Deviation	0.017	0.003	0.060	0.007	0.011
Correlation with Y	1	0.792	0.989	0.898	0.964

Table 20.10 lists the business cycle statistics. The results are very similar to Table 20.8 in the previous section. The exception is that the standard deviation of L is one order of magnitude larger, reflecting the larger variability of unemployment.

20.9 Heterogeneity of jobs and the frictional wage dispersion

So far, we have assumed that jobs are homogeneous, and workers always accept an offered job. In this section, we introduce a model where job offers are heterogeneous. Some jobs pay more than other jobs, and the kind of jobs offered to the worker is stochastic. It is assumed that every period, an unemployed worker can receive only one job offer. After receiving the offer, she decides whether to accept it. The model in this section is called the McCall search

model (McCall, 1970) or simply the search model. The search model focuses on the worker's decision, and the demand side is simplified. Because the worker's decisions are operative margin, labor supply plays an active role.

The heterogeneity of job offers gives rise to wage dispersion. The labor market friction plays a crucial role; if there are no frictions, all workers will accept only the best (highest-paying) job. Even with labor market frictions, it is not trivial to think of a setting where firms actively offer different wage levels. A well-known example is called the *Diamond paradox*. Diamond (1971) has shown that, if the jobs are homogeneous and the worker has to leave the job to look for another job, all firms offer the workers' *reservation wages* (the lowest wage the worker would accept) even with a small search cost. It is easy to check that this outcome constitutes a Nash equilibrium: workers do not look for another job if they know all other firms are offering their reservation wage, and no firm would want to deviate. The Nash equilibrium is unique because, with any other wage offer distribution, the firm that offers the highest wage has an incentive to lower the wage slightly. In this section, instead of considering firms' wage setting behavior explicitly, we assume away the heterogeneity of wage offers. This type of firm behavior can be justified by relaxing Diamond's assumptions. For example, one can assume that the jobs are heterogeneous or workers can search on the job.

Formally, suppose that the worker is infinitely lived and the utility of the worker is linear:

$$\mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t c_t \right].$$

An unemployed worker earns $b > 0$ every period. This income can be interpreted as home production, unemployment insurance benefit, or the value of leisure. Each worker receives one job offer every period. Job offers differ in terms of the wage w. It is stochastic with distribution function $F(w)$, which has a lower bound of 0 and upper bound w^u. The Bellman equation for an unemployed worker is therefore written as

$$U = b + \beta \int_0^{w^u} \max\{W(w), U\} dF(w), \tag{20.52}$$

where U is the value of unemployment and $W(w)$ is the value of employment with wage w. The expression (20.52) can be rewritten as

$$(1-\beta)U = b + \beta \int_0^{w^u} \max\{W(w) - U, 0\} dF(w). \tag{20.53}$$

Every period, an employed worker faces the probability $\sigma \in (0,1)$ of losing her job. Therefore, the Bellman equation for an employed worker is

$$W(w) = w + \beta[(1-\sigma)W(w) + \sigma U]. \tag{20.54}$$

Equation (20.54) can be rewritten as

$$W(w) = \frac{w + \beta\sigma U}{1 - \beta(1-\sigma)}. \tag{20.55}$$

From this expression, we can see that $W(w)$ is increasing in w and $W(0) = \beta\sigma U/(1-\beta(1-\sigma)) < U$. We assume that w^u is sufficiently large so that $W(w^u) > U$. Therefore, there exists a threshold w^* where $W(w^*) = U$, $W(w) > U$ for $w > w^*$, and $W(w) < U$ for $w < w^*$. In other words, the wage level w^* is the worker's *reservation wage*. The value of w^* characterizes the worker's choice in this model.

Plugging the expression (20.55) into (20.53) and using that $W(w) > U$ if and only if $w > w^*$,

$$(1-\beta)U = b + \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [w - (1-\beta)U] dF(w). \tag{20.56}$$

Considering the expression (20.55) for $w = w^*$ and noting $W(w^*) = U$, we obtain

$$(1-\beta)U = w^*.$$

Thus (20.56) can be rewritten as

$$w^* = b + \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [w - w^*] dF(w). \tag{20.57}$$

This equation solves the reservation wage w^*.

What can we learn from this model? First, let us consider the frequency of job acceptance. The worker only accepts the jobs that are better than w^*. Thus the *job finding probability* λ is

$$\lambda = 1 - F(w^*).$$

The job finding probability is decreasing in w^*: when the workers are choosier, they find jobs less often.

We can also conduct various comparative statics to analyze how changes in parameters affect the reservation wage w^* (and therefore λ). Rewrite (20.57) as:

$$\mathbf{G}(w^*, \beta, \sigma, b) = 0,$$

where

$$\mathbf{G}(w^*, \beta, \sigma, b) \equiv w^* - b - \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [w - w^*] dF(w).$$

It is straightforward to show that $\partial\mathbf{G}/\partial\beta < 0$, $\partial\mathbf{G}/\partial\sigma > 0$, and $\partial\mathbf{G}/\partial b < 0$. For w^*, because (using Leibnitz's rule)

$$\frac{\partial}{\partial w^*} \int_{w^*}^{w^u} [w - w^*] dF(w) = -[1 - F(w^*)],$$

$\mathbf{G}(w^*, \beta, \sigma, b) > 0$. From the implicit function theorem, w^* is increasing in β and b and decreasing in σ. Intuitively, the worker becomes choosier (w^* becomes larger) when β increases because the future gain from a better job has a higher weight compared to the opportunity loss from missing the immediate job. An increase in b makes the unemployment state more attractive and induces workers to wait longer. A higher σ implies that even a good job won't last long, and thus it becomes less attractive to wait for a good job offer to arrive.

An interesting comparative-statics exercise with this class of model is to analyze the effect of changes in the wage offer distribution. First, consider the change in the average wage. To analyze the change in average, suppose that the wage offer is $w+\varepsilon$ instead of w above (with the same distribution for w), and how the change in ε changes the reservation wage $w^*+\varepsilon$ when evaluated at $\varepsilon = 0$. Now the $\mathbf{G}$ function is modified to

$$\mathbf{G}(w^*, \varepsilon) \equiv w^* + \varepsilon - b - \frac{\beta}{1-\beta(1-\sigma)} \int_{w^*}^{w^u} [(w+\varepsilon) - (w^*+\varepsilon)] dF(w).$$

$$\frac{\partial}{\partial \varepsilon} \mathbf{G}(w^*, \varepsilon) = 1$$

and (using Leibniz's rule, evaluated at $\varepsilon = 0$)

$$\frac{\partial}{\partial w^*} \mathbf{G}(w^*, \varepsilon) = 1 + \frac{\beta}{1-\beta(1-\sigma)} [1 - F(w^*)].$$

Thus

$$\frac{dw^*}{d\varepsilon} = -\frac{1-\beta(1-\sigma)}{1-\beta(1-\sigma)+\beta[1-F(w^*)]}.$$

The change in the reservation wage is, therefore,

$$\frac{d(w^*+\varepsilon)}{d\varepsilon} = \frac{dw^*}{d\varepsilon} + 1 = \frac{\beta[1-F(w^*)]}{1-\beta(1-\sigma)+\beta[1-F(w^*)]} \in (0,1).$$

Thus the reservation wage goes up, but not one-to-one. When the average wage offer goes up by one dollar, the reservation wage goes up by less than one dollar. This outcome arises because b is kept constant. Because b is the same, the relative attractiveness of the unemployment state (compared to working) goes down. This effect attenuates the effect of the change in the wage offer distribution.

Next, consider the dispersion of the wage offer distribution. To analyze the effect of dispersion, we first have to define the appropriate concept of the dispersion of wage offers in this context. Here, we introduce the concept of the *mean-preserving spread*. For a random variable x with the distribution function $F(x)$, we can construct a random variable $\tilde{x} \equiv x+z$ where z has a distribution function $H_x(z)$ and its mean is zero ($\int z dH_x(z) = 0$). Then, the mean of $x+z$ is x, and let us call the new distribution function $G(\tilde{x})$. Then we refer to $G(\cdot)$ as a mean-preserving spread of $F(\cdot)$. It can be shown (see Mas-Colell, Whinston, and Green, 1995, p. 198) that $G(\cdot)$ being a mean-preserving spread of $F(\cdot)$ is equivalent to

$$\int_0^x G(t)dt \geq \int_0^x F(t)dt \text{ for all } x. \tag{20.58}$$

Now let us rewrite the equation (20.57) as[14]

$$\begin{aligned} w^* &= b + \frac{\beta}{1-\beta(1-\sigma)} \left[\int_0^{w^u} [w-w^*]dF(w) - \int_0^{w^*} [w-w^*]dF(w) \right] \\ &= b + \frac{\beta}{1-\beta(1-\sigma)} \left[\mu_w - w^* - \int_0^{w^*} [w-w^*]dF(w) \right], \end{aligned} \tag{20.59}$$

[14]This derivation follows Ljungqvist and Sargent (2012, pp. 166-167).

where μ_w is the mean value of w. Integration by parts yields

$$\int_0^{w^*} [w - w^*] dF(w) = -\int_0^{w^*} F(w) dw$$

and thus (20.59) can be rewritten as

$$w^* - b - \frac{\beta}{1 - \beta(1 - \sigma)} \left[\mu_w - w^* + \int_0^{w^*} F(w) dw \right] = 0. \tag{20.60}$$

It is straightforward to show that the left-hand side is increasing in w^*. Suppose that the distribution of w, $F(w)$, becomes more dispersed in the sense of the mean-preserving spread. The property (20.58) implies that the left-hand side of (20.60) goes down with this change in the distribution, and thus w^* has to go up. The reservation wage increases with the dispersion of the wage offers. Intuitively, the worker becomes choosier with more dispersed wage offers because the possibility of a good wage offer increases. With higher dispersion, the possibility of a bad offer also increases, but the left tail of the distribution does not matter because these offers are rejected in any case. In other words, the option value of searching increases with the dispersion of the wage offer.

Now, consider the model implications for the realized wage dispersion. The equilibrium wage dispersion in this model is often called the *frictional wage dispersion* because all workers would work at $w = w^u$ if there are no search frictions (i.e., if all jobs are available to the workers). Workers accept a job with $w < w^u$ because it is costly to wait for high-wage job offers. To analyze the frictional wage dispersion, first define the mean (accepted) wage as

$$w^M \equiv \frac{\int_{w^*}^{w^u} w dF(w)}{1 - F(w^*)}.$$

Let us also define

$$\rho \equiv \frac{b}{w^M},$$

that is, the ratio of the unemployed worker's income to the mean wage. When b is interpreted as the income from unemployment insurance, ρ corresponds to the replacement rate. Further, as defined above, let $\lambda \equiv 1 - F(w^*)$ be the job-finding probability, that is, the probability that an unemployed worker transitions into employment.

Define the *mean-min ratio* of the (accepted) wages, Mm, as

$$Mm \equiv \frac{w^M}{w^*}.$$

Note that w^* is the minimum value of the accepted wages. Then, using (20.57) and the above definitions, we can derive

$$Mm = \frac{1 + \beta\lambda/(1 - \beta(1 - \sigma))}{\rho + \beta\lambda/(1 - \beta(1 - \sigma))}.$$

A back-of-the-envelope calculation with (monthly) $\beta = 0.996$, $\sigma = 0.034$, $\lambda = 0.45$, and $nb = 0.4$ yields $Mm = 1.031$. The mean wage is only 3.1% larger than the lowest wage in

this economy. In other words, search friction can explain only a tiny part of the observed wage dispersion in this model. This result is often referred to as the *frictional wage dispersion puzzle* in the literature. The reason is that, in the model with this parameterization, it is not very costly to wait for a new job (the offer comes fairly frequently), whereas the benefit of receiving a better wage offer is large. One situation where the frictional wage dispersion is high is, therefore, when the cost of unemployment is high. A large cost of unemployment makes unemployed workers accept low wage offers to avoid unemployment. Another situation is when workers can search on the job. For example, if the offered wage distributions are identical and the offer frequency is the same for on- and off-the-job search, the worker would accept any job with $w > b$ when unemployed. In this case, the worker does not have to give up the option value of search when accepting a wage offer.

Chapter 21

Inequality

Per Krusell and Víctor Ríos-Rull

21.1 Introduction

Traditionally, inequality has simply not been a topic of its own within macroeconomics. Arguably, it has been present indirectly in traditional textbooks through the focus on unemployment, but then mostly as an indicator of how the degree of inefficiency of the aggregate economy moves up and down with the business cycle. Today, the situation is quite different. Since at least the turn of the millenium, inequality has risen and remained high on the macroeconomic research agenda, and macroeconomic policymakers, including central banks, pay significant attention to it.

There are several reasons for the current focus on inequality. One is that many, if not most, economists view significant inequality as a potential cause of concern, especially since many measures reveal a persistent and still ongoing increase in the concentration of earnings and wealth, since the last decades of the past century. At the same time, we recognize that inequality in many ways is a natural result of the working of an efficient market economy. Thus, at the very least we need to understand what explains the observed trends. Second, and relatedly, in many ways the determination of inequality is, by its nature, a macroeconomic phenomenon, i.e., one where general equilibrium interactions are important. For example, as we shall see, modern macroeconomic modeling naturally gives rise to theories of both wage and wealth inequality.

A third, and quite independent, reason for studying inequality in macroeconomics is that there is increasingly convincing evidence that it affects aggregates. How, for example, fiscal and monetary policy changes propagate through the economy critically hinges on the marginal propensities to consume, invest, and work throughout the population; this was evidenced in Chapter 11, where we learned that the propensity to consume varies in the full range between close to zero, for many consumers, and near one, for another non-trivial part of the population of households. A main reason behind the interest in heterogeneous-agent models is precisely that they admit this kind of heterogeneity in propensities. Moreover, they can be parameterized to match microeconomic data so as to maintain quantitative discipline when conducting counterfactual experiments. As a result, this new literature holds, the new generation of macroeconomic models can produce robust predictions and be very valuable for policymakers.

The goal of the present chapter is to briefly review some key data and then go over some of the main ways in which macroeconomic theory has addressed inequality. It is not a survey and therefore aims to keep references at a minimum and instead lay out the key facts and theories in a relatively compact way. It also omits some topics altogether. For example, how inequality influences aggregates through the political system is not addressed, even though it is arguably very important. There are two theory sections. The first one, Section 21.2, addresses the determinants of inequality: what macroeconomic theory predicts. In this section, the focus is first on the *skill premium*, i.e., channels through which highly educated workers earn more than the less educated. Then it presents the core macroeconomic theories of wealth (and consumption) inequality, what is broadly known as *heterogeneous-agent* models. Thereafter, Section 21.3 discusses how the presence of inequality affects the workings of macroeconomic aggregates: why inequality matters for macro.

21.1.1 Data

The idea in this section is to very briefly go over some evidence on inequality measures for the microeconomic equivalents of macroeconomic variables: labor income, wealth, hours worked, and consumption. As in most of the book, the focus is on the U.S., so we encourage the reader to search for the equivalent information for other countries. We begin with static descriptions and then describe some changes over time.

The cross-section

Beginning with the income distribution for U.S. households in 2022, consider Figure 21.1. It shows a histogram (the height of each equal-sized interval represents the fraction of households is such an interval). Its main feature is that it is *highly skewed*: a shape that rises steeply at 0, peaks below $50,000, and then slowly decreases, with a thick right tail. The right tail keeps increasing far beyond the maximum value on the x-axis ($800,000) and still has non-negligible mass at levels 1,000 times that. As a result of the significant skewness, mean income is far higher than median income, a feature that holds in all countries for which there is reliable data.

The same qualitative skewness can be recorded for labor earnings, as well as for wealth. To compare these distributions, it is useful to use a Lorenz curve and, based on it, compute Gini coefficients. A stylized Lorenz curve is depicted as the solid, convex curve for income in Figure 21.2. A point (x, y) on the curve describes the share y of overall income earned by the poorest x percent of the population. Perfect income equality would mean that the Lorenz curve is the 45-degree line (plotted as a dashed line). The Lorenz curve is always below the 45-degree line by construction, as the population is ranked in incomes from left to right. The most extreme income inequality would be a Lorenz curve that is a flat line (equal to the x-axis) up until 100, where it jumps to 100. The Gini coefficient is computed as the area A divided by the area A+B, i.e., the size of the full triangle under the 45-degree line. When there is perfect equality, the Gini coefficient is therefore 0; a Gini coefficient of 1 is the extreme opposite.

Turning now to the U.S. Lorenz curves, Figure 21.3 shows, simultaneously, the wealth and labor income distributions. The former is displayed for net wealth (which includes all

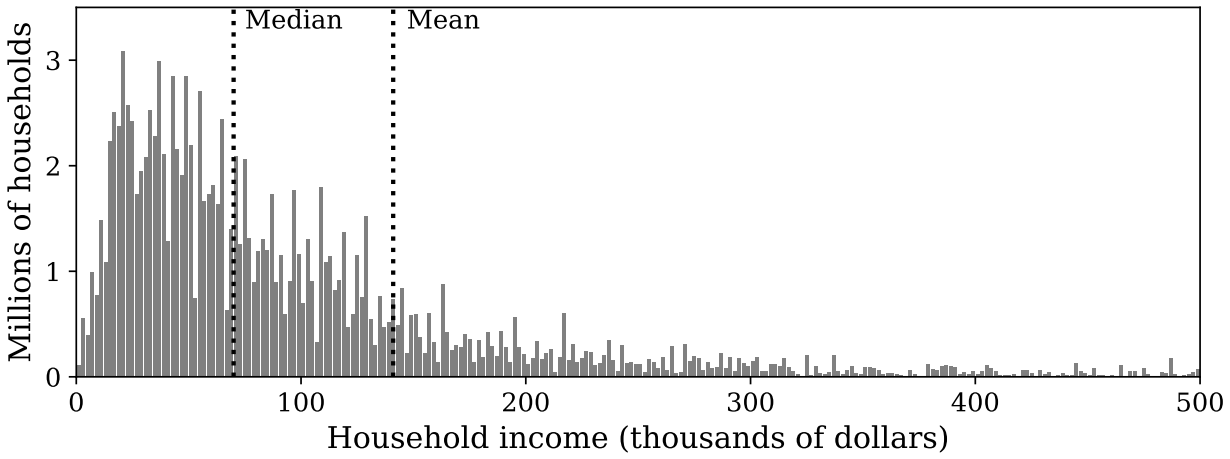


Figure 21.1: Histogram of the income distribution.

Source: Kuhn and Rios-Rull (2025) using the 2022 Survey of Consumer Finances.

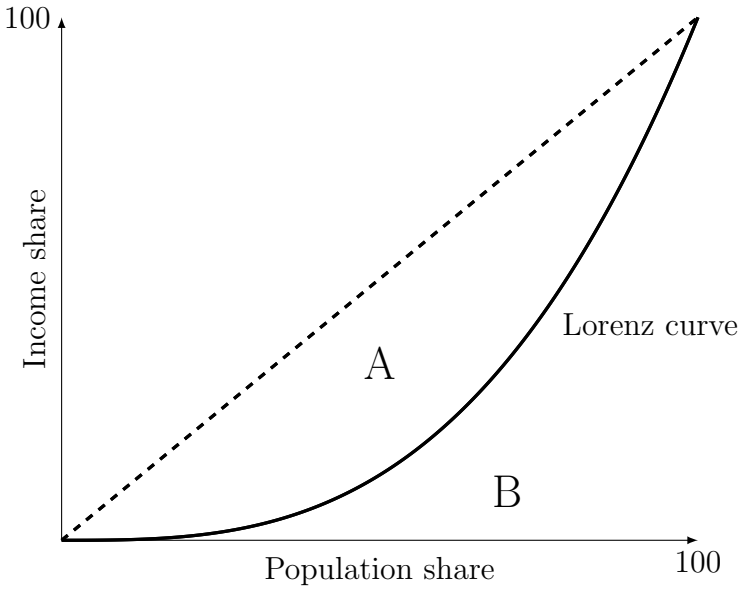


Figure 21.2: Lorenz curve.

assets net of debts) as well as for residential wealth only, and the latter in two versions as well (labor income includes a share of self-employed business income, whereas wage income simply has salaries). The Gini coefficients for each curve are displayed in the legend. We see, first, that the distribution for net worth, with a Gini at 0.83, is far more dispersed than those for the other distributions. Residential wealth is the least dispersed, with a Gini of 0.66, and the two labor income distributions are just slightly more dispersed, with Ginis around 0.68. The finding that wealth is more dispersed than labor income holds qualitatively over time and for all other economies for which we have seen studies.

Table 21.1 displays distributional information for earnings, total income, wealth, consumption, and hours worked. We learn that the earnings- and wealth-poorest 10 percent of

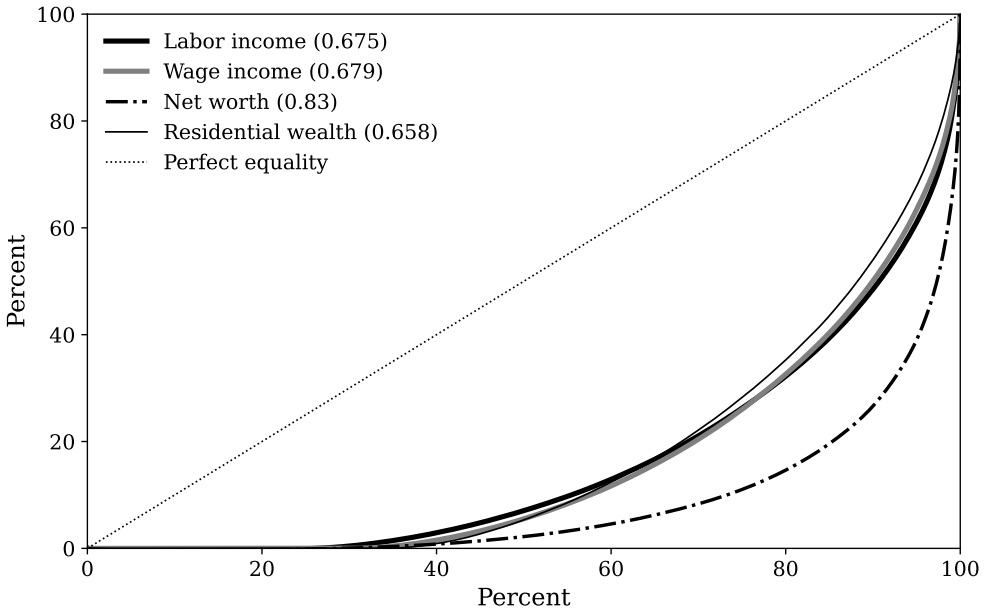


Figure 21.3: Lorenz curves of labor earnings, residential wealth, and net worth.

Source: SCF 2022. See Kuhn and Rios-Rull (2025).

the population of households have no earnings or wealth, respectively, while the 1-percent earnings- and wealth-richest have 19.4 and 35.1 percent of the total earnings and wealth, respectively. Especially the 35.1 number is staggering: over 1/3 of all wealth is held by the 1-percent wealthiest. The measurement of wealth is difficult, as there is no registry data on wealth and surveys offer a very partial account and have trouble sampling the very richest; for the richest groups, off-shore (unrecorded) wealth is another challenge. Furthermore, the taxation of wealth (either directly or via estate taxes) gives incentives to under-report the value of assets, especially for those businesses that are not publicly traded. One way to estimate wealth is from tax returns, by looking at capital income and inferring the stock from the income flow. Yet, in countries where wealth is taxed, registry data shows similar qualitative features to those in the table for the U.S.

Another characteristic often used in characterizing income and wealth distributions is the approximate fact that the right tails of the distributions are Pareto-shaped. A Pareto distribution for a variable x is characterized by a linear relation between the logarithm of x and the logarithm of one minus the cumulative distribution at x. The slope of this relationship is negative and defines, by its inverse, the thickness of the tail.[1] The wealth distribution has a significantly thicker right tail than the earnings distribution.

Table 21.1 also displays information on consumption and hours worked.

First of all, consumption is much less dispersed than earnings, income, and wealth.[2] The

[1]The cumulative function is $1-(x/\underline{x})^{\alpha}$, where $x \geq \underline{x}$, the minimum value for x, and $-\alpha$ is the slope referred to.

[2]This is in part a measurement issue: the distribution is trimmed at the top and the bottom percentiles,

Table 21.1: 2022 Per household shares of selected groups sorted by each variable.

	Bottom			Quintiles					Top		
	0-1	1-5	5-10	0-20	20-40	40-60	60-80	80-100	90-95	95-99	99-100
Earnings	−0.00	0.00	0.00	−0.16	0.50	10	0.96	3.39	2.5	4.88	19.4
Income	0.00	0.08	0.12	0.14	0.30	0.50	0.82	3.24	2.5	4.45	22.4
Wealth	−0.2	−0.02	0	−0.01	0.05	0.19	0.50	4.70	2.48	6.48	35.1
Consumption				0.44	0.66	0.84	1.12	1.93			
Hours worked											
per household	0.07	1.05	2.25	9.31	12.64	16.84	23.54	37.66	8.18	8.04	7.65
per person	0.09	1.28	2.65	11.65	19.89	20.57	20.97	26.91	6.67	6.27	2.07

Note: Shares of the earnings, income, and wealth distribution in 2022. For earnings, income and wealth the source is Kuhn and Rios-Rull (2025). For consumption data from Table 1101, https://www.bls.gov/cex/tables/calendar-year/aggregate-group-share/cu-income-quintiles-before-taxes-2022.pdf.

low dispersion compared to wealth is also to be expected from standard permanent-income theory: consumption equals the flow equivalent of present-value earnings, which is the largest part for most people, plus the return on wealth. In addition, private insurance arrangements, as proposed by Krueger and Perri (2006), that are not measured in the data will also make consumption dispersion lower.

Hours worked are also much less dispersed than are earnings or wealth, especially as measured per person. Though we see that some people work many more hours than others (e.g., the 1-percent hardest working contribute almost 8 percent of the total working time), trying to see systematic patterns as to who works more and who works less is much more difficult. Figure 21.4 shows, for example, that across wage bins, the hours of work for men are very evenly distributed; these are *residualized* hours worked, i.e., some observables have been controlled for, namely, age, age squared, education, and race, as also discussed in Chapter 12.[3] That is, it is not that the most productive work significantly more, at least not if productivity is well approximated by wages. The figure also shows that the last two decades have seen a marked fall in the number of hours worked for the wage-poorest.

Given the limited dispersion in hours, and the fact that wages and hours are not strongly correlated, it follows—given the highly dispersed earnings—that wages exhibit significant dispersion. When discussing trends below, we will look at one measure of wage dispersion, namely, that between workers with different educational degrees, and its evolution over time. Wages of course differ within educational groups too, in part due to experience and the worker's age; in addition, there are observable characteristics like gender and race that also systematically influence wages. Whether the latter factors are due to discrimination of some sort is an important issue from a macroeconomic perspective: a society that does not allow all its individuals to flourish will under-perform in terms of efficiency. However, we do not discuss it further in this chapter.

as the Consumer Expenditure Survey does not include reliable measures at the extremes.

[3]The corresponding graph for women is very similar.

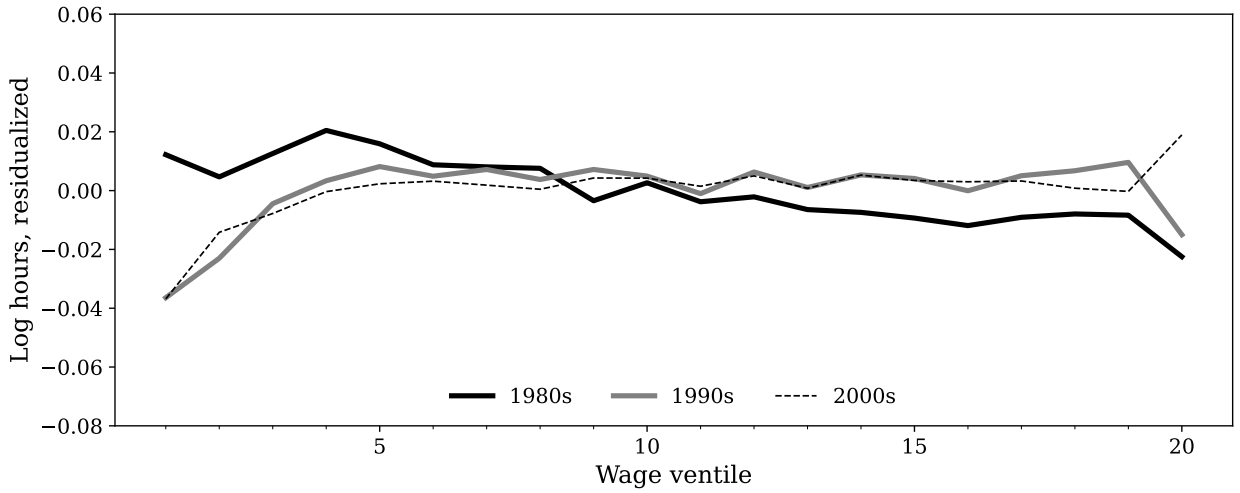


Figure 21.4: (Log) hours worked by wage ventiles.

Note: Ventiles are 5% bins. **Source**: Boppart et al. (2024).

We now look at the returns to capital—the per-dollar returns on the savings of households—and how they are dispersed in the population. We know that different assets give different returns and that these returns depend on their risk characteristics, as discussed in Chapter 16; for example, highly liquid savings give much lower average return than stock. Moreover, housing or land wealth delivers returns that are not financial—the services enjoyed by those who live in the house or use the land—but house and land prices also fluctuate significantly across locations and time.

At least for financial assets, one might have expected—from basic principles of portfolio management—that all consumers hold the same portfolios and enjoy the same returns on their portfolios, i.e., the return on "the market portfolio." Until relatively recently, little data was available on the portfolio holdings of individuals across the population. Fortunately, however, we now have data from countries where registry data (i.e., data for the whole population) is available, and increased efforts have also been made to unveil information implicit in capital incomes and from surveys. Thus, we now know more, and we know better: using Swedish registry data, Bach, Calvet, and Sodini (2020) report that the portfolio shares vary greatly and systematically across the wealth dimension. The wealth-poorest all have mostly cash (defined to include deposits) but from the 50th until the 95th percentile, residential real estate is the biggest component. The further to the right in the distribution of wealth, the higher the share of risky financial wealth and, at the very top, private equity.[4]

The portfolio shares would perhaps be uninteresting if it were not for the fact that the different portfolios have different return characteristics. In Figure 21.5, which is a time-series average borrowed from Hubmer, Krusell, and Smith (2018), we display these return characteristics for the same wealth percentiles.[5]

[4]Private equity here means equity that is not publicly traded.

[5]For any cell, take the portfolio shares from Figure 2 of Bach et al. (2020) and apply a return to each component. The component derived from U.S. data, when available, and otherwise for the Swedish data from Bach et al. (2020). The U.S. data sources are Kartashova (2014), for public and private equity, Jordà, Knoll, Kuvshinov, Schularick, and Taylor (2019), for bonds, and the real estate return is based on the Case-Shiller

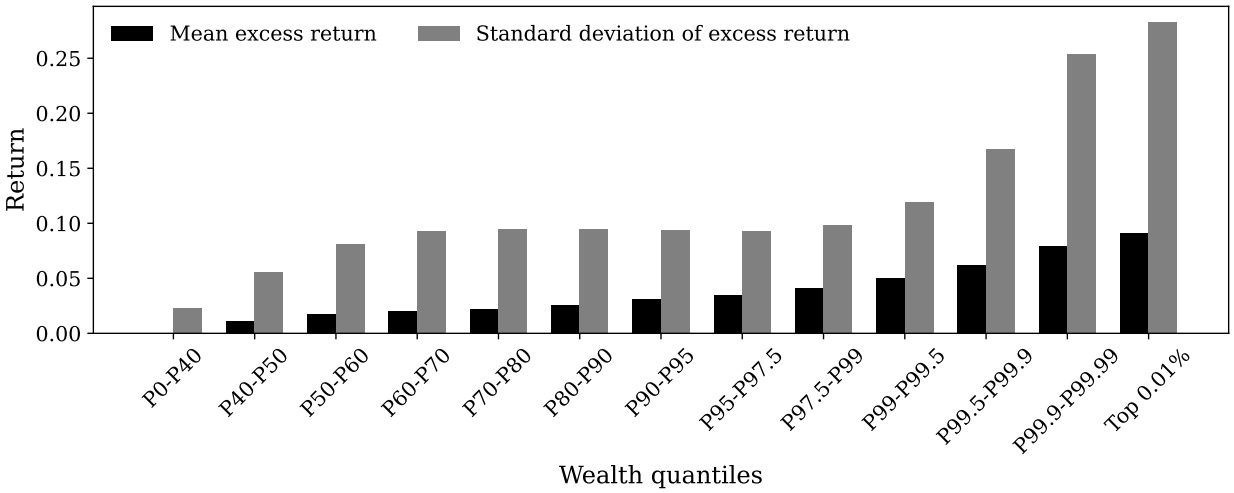


Figure 21.5: Mean and standard deviation for excess returns on portfolios.

The figure reveals strikingly large differences in mean returns—the darker bars—by wealth percentile: the higher up you are in the wealth rank, the higher is your return. Part of this comovement can of course be due to high returns causing high wealth, but an important part is surely just due to the systematic differences in returns across assets. Secondly, note that cash and deposits provide liquidity services by facilitating payments and these benefits are not captured in returns. Also, recall that the returns on housing do not include the housing service associated with owner occupancy. A second important feature of the figure is the significant standard deviation in returns and how, at the very top of wealth the distribution, the standard deviation shoots up significantly. Extremely high return outcomes are a powerful source of wealth build-up, and we will discuss this mechanism in the theory section below.

Trends

We first look at the movements of capital and wealth relative to GDP (or income); these are relevant as capital and wealth are so concentrated among the wealthiest, whereas the large bulk of the population mostly rely on their flow income. Thus, Figure 21.6 shows that although the capital-GDP ratio has been very stable at around 3.5, there has been a marked upward trend in the wealth-to-income ratio, by around 50%. Note that wealth includes assets beyond capital, such as land and claims on the government.[6]

Turning to the inequality within earnings and wealth, Figure 21.7 contains two panels and allows us to make a number of observations.

While average real earnings have risen—the solid line in the left panel of the figure—we see that the median and the 30th percentile have seen no real earnings growth since 1990. Thus, the top earners have had significant earnings growth, as evidenced by the 90th

index.

[6]The figure shows the data from two sources, yielding a similar qualitative message but a slightly higher growth rate from the SCF survey measure.

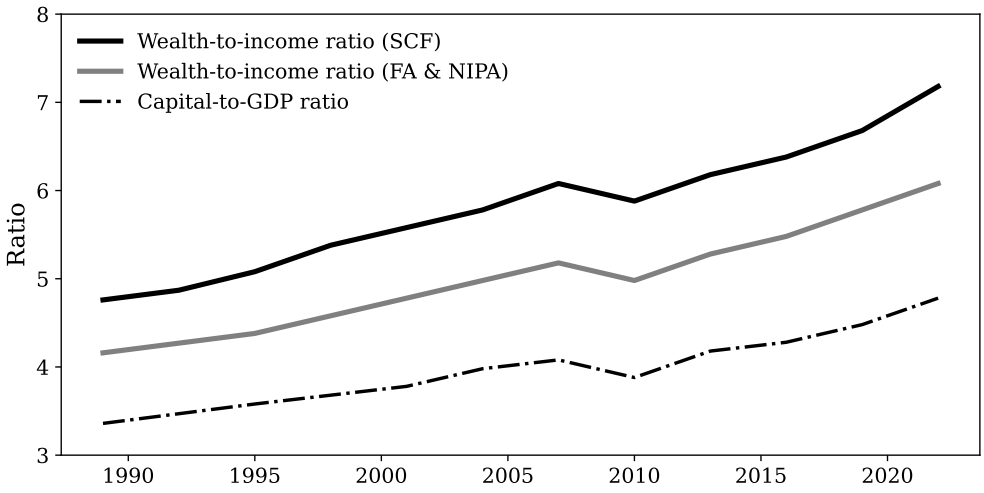


Figure 21.6: Wealth-to-income and capital stock-to-GDP ratios by SCF wave.

Note: Wealth-to-income ratio based on the SCF, wealth-to-income ratio using income from National Income and Product Accounts (NIPA), and wealth from Financial Accounts (FA), and capital stock-to-GDP ratio from Penn World Table 10.01. **Source**: Kuhn and Rios-Rull (2025).

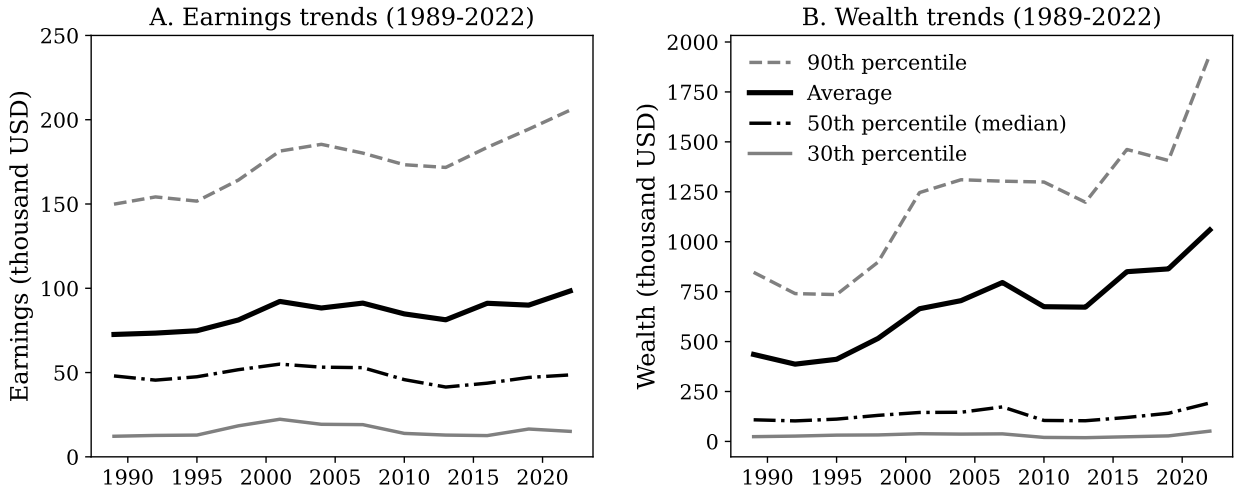


Figure 21.7: Inequality statistics for the evolution of earnings and wealth 1989–2022.

Note: These figures show the evolution of earnings and wealth across different percentiles of the distribution using data from the Survey of Consumer Finances (SCF). Panel (A) shows earnings trends and panel (B) shows wealth trends. All values are in thousands of 2022 USD.

percentile earnings growth rising from 150,000 to over 200,000 over this period (in 2022 USD). That is, earnings inequality has grown. We will show one aspect of this increase in dispersion, the rising skill premium, just below.

In the right panel of Figure 21.7, we see that real wealth has more than doubled on average over the same time period, but again with very limited changes for the median or in the lower percentiles: the wealth buildup has occurred at the top, from around 800,000 to almost 2,000,000 (in 2022 USD). In sum, both earnings and wealth inequality have seen increases, mainly through rapid growth at the top of the distribution and none at the bottom.

Returning to the trends in earnings inequality, let us begin with trends in hours. Overall, hours are stable in the U.S. over the postwar period, but there are movements under the surface. An important determinant of hours worked is the extent of unemployment; data on that, revealing no long-run trend but large fluctuations over the business cycle, was discussed in Chapter 20. Another component is labor-force participation, and we displayed data on that in Chapter 12 and documented, in particular, that women's employment rate has risen dramatically. Today female labor force participation is only 10 percentage points below the value for males; in the middle of the 20th century, the difference was 50 percentage points. These factors, however, are minor in importance in comparison with the wage paid (on average) per hour: these differ drastically across the population and is the main reason why earnings are so highly dispersed. Figure 21.8 shows the development of the skill premium—the wage gap between those with a college degree and those without—over time.

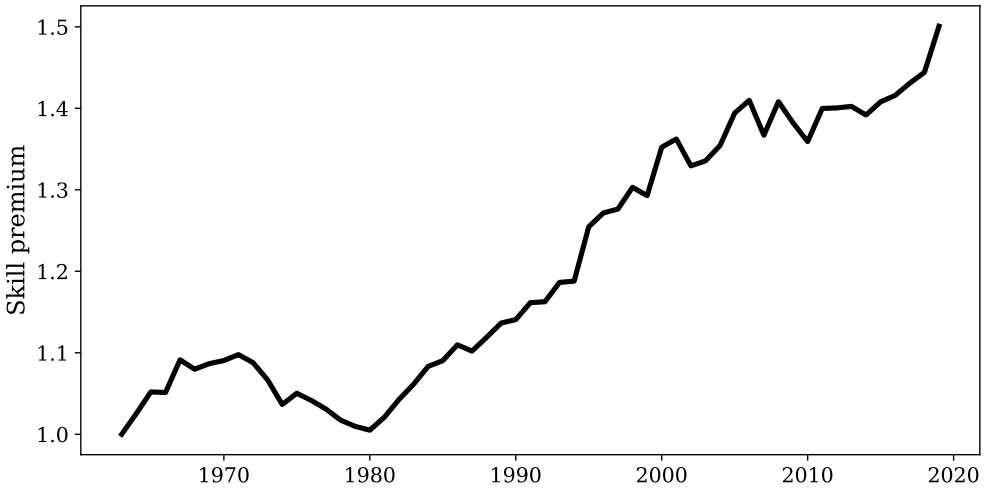


Figure 21.8: Evolution of the skill premium over time.

Source: Ohanian et al. (2023).

In 1963, the premium was around 50%. The time series in the figure, normalized to be 1 in 1963, shows the growth since then. We see that the premium has risen by 50%, rather steadily since 1980, after a dip during the 1970s. This increase is particularly striking given that the number of college graduates has risen at a high rate during a number of decades during the second half of the twentieth century (but less so in recent decades). Of course there is also wide dispersion within the group with a college degree and within the one without; these dispersions have also increased over time.[7]. A source of wage dispersion that has gone the other way is the wage wedge between men and women (for otherwise

[7]A graduate student may be especially interested in learning that the graduate school premium also has risen; for an early study, see Eckstein and Nagypal (2004)

observationally equivalent workers); it is now less than 10 percent.[8]

21.2 Theory: macroeconomics and inequality

The present section and the next sections address the facts in the previous section using macroeconomic models. We first look at macroeconomic factors that determine important parts of the observed inequality. In particular, macroeconomic models are useful because the extent and shape of inequality are often affected by general-equilibrium interactions. Since the macroeconomic models we use have explicit microeconomic foundations, they also allow us to neatly separate partial- from general-equilibrium effects. Second, we look at how the presence of, and possible movements in, inequality affects macroeconomic aggregates.

We begin the present section by addressing some basic income inequality facts and then move to wage inequality. We finally discuss wealth inequality and, in that context, briefly also touch on inequality in hours worked.

21.2.1 The labor share and the capital-output ratio

From the perspective of the labor share mainly going to "workers" and the capital share to "capitalists", or "rentiers", it becomes relevant to study the behavior of these shares, as well as of the ratio of capital to output.

A first observation is then that our basic, neoclassical models have implications for the labor share of income. The most common assumption in the applied literature is that the aggregate production function is of the Cobb-Douglas variety, $A_t k_t^{\alpha} h_t^{1-\alpha}$. Here, TFP moves over time but the output elasticities with respect to capital and labor are constant, α and $1-\alpha$, respectively, and will equal the corresponding income shares under perfect competition. Therefore, the income shares are time-invariant, also off the balanced growth path. We saw, however, in Chapter 2 that the observed labor share has been declining in many countries in recent times. Can such a development occur in the neoclassical model?

Let us examine the labor share for a more general, constant-returns-to-scale production function $F(A_{k,t}k_t, A_{h,t}h_t)$, where the shape of F is time-invariant and technological change occurs via the capital- and labor-augmenting factors A_k and A_h. Under the assumption of perfectly competitive input markets, the share becomes

$$\frac{r_t k_t}{y_t} = \frac{A_{k,t} F_1\left(A_{k,t}k_t, A_{h,t}h_t\right) k_t}{F\left(A_{k,t}k_t, A_{h,t}h_t\right)} = \frac{F_1\left(1, \frac{A_{h,t}h_t}{A_{k,t}k_t}\right)}{F\left(1, \frac{A_{h,t}h_t}{A_{k,t}k_t}\right)},$$

where $F_i(\cdot,\cdot)$ is the partial derivative with respect to the ith argument, and we have used the homogeneity of degree 1 of F. Clearly, the share only depends on $(A_{h,t}h_t)/(A_{k,t}k_t)$, which is constant on a balanced growth path but off the balanced path it is changing and how the share then moves depends on the shape of F. Assume that it is a CES function, with

[8] See, e.g., Blau and Kahn (2017). Men are still vastly over-represented among the very highest earners.

elasticity ρ.[9] Then

$$\frac{r_t k_t}{y_t} = \frac{1}{1 + \left(\frac{1-\alpha}{\alpha}\right)^{\frac{1}{\rho}} \left(\frac{A_{h,t} h_t}{A_{k,t} k_t}\right)^{\frac{\rho-1}{\rho}}}.$$

We see that if $\rho > 1$, i.e., under stronger substitutability than Cobb-Douglas, a decrease in $(A_{h,t}h_t)/(A_{k,t}k_t)$ increases the capital share. A decrease in $(A_{h,t}h_t)/(A_{k,t}k_t)$, furthermore, is generated by the capital input growing faster than the labor input, which in turn results if there is investment-specific technological change, as described in Chapter 13. Hence, investment-specific technological change can cause a lower labor share; Karabarbounis and Neiman (2014) advance this theory, with an estimate of ρ slightly above 1 along with a decline in the relative price of capital.

The labor share can also be affected by departures from perfect competition, because firms' product-market power changes, affecting firms' margins,/ or because the power has shifted between employers and employees (monopsony vs. union power, respectively). We do not provide an analysis of these mechanisms here but they have attracted significant research over the last decade.[10]

Turning to k/y, the value of the stock of capital relative to output, we note that along a balanced growth path,

$$sy_t = (\delta + g)k_t \quad \Rightarrow \quad \frac{k_t}{y_t} = \frac{s}{\delta + g},$$

where g is the (labor-augmenting) growth rate, s is the saving rate, and δ the depreciation rate. Thus, a higher long-run growth rate makes k/y fall, but g is small relative to δ, at least based on historically recorded levels.[11] As for transitional dynamics, $k/y = k/F(k,h)$ is an increasing function of k, so k/y will tend to increase (decrease) toward its steady-state value if k is initially below (above) its steady-state value.

The k/y ratio can also be thought of as the wealth-to-output ratio in the standard neoclassical model, at least when wealth is thought of as an asset: the asset is k.[12] As we saw above, in the data, assets, broadly defined, have several components: land (a very large component), liquid, low-return assets (narrowly defined as cash and liquid deposits), bonds (some risky, some less so), and stock (both publicly traded and private).[13] As we saw in our introductory Chapter 2, the k/y is roughly 3 for annual data, but assets more broadly defined is at least twice that. From the perspective of inequality, the broader measure is interesting for several reasons. First, some of the additional assets are concentrated particularly among the wealthiest and therefore are very important for a full assessment of inequality. Second and relatedly, the prices, as well as the returns, on some of these assets are highly variable over time. Third, from our perspective here, general-equilibrium modeling of the broadest

[9] That is, we have $F(x,y) = \left(\alpha^{\frac{1}{\rho}} x^{\frac{\rho-1}{\rho}} + (1-\alpha)^{\frac{1}{\rho}} y^{\frac{\rho-1}{\rho}}\right)^{\frac{\rho}{\rho-1}}$.

[10] See, e.g., De Loecker, Eeckhout, and Unger (2020).

[11] Piketty (2014) formulates a so-called "Second Fundamental Law of Capitalism" where, based on a non-standard version of Solow's growth model where as g goes to 0, $k/(y - \delta k)$ goes to infinity. The non-standard aspect is in the formulation of saving: the saving rate is constant in net terms, i.e., net investment (the increase in the capital stock) is a constant fraction of net output (output net of capital depreciation). See Krusell and Smith (2014) for details.

[12] One can alternatively define wealth to incorporate human wealth, present and future.

[13] When computing net assets, of course, various forms of debt, such as mortgage loans, must be subtracted.

view on assets requires a model that goes beyond the simple neoclassical model: (i) it needs land and housing modeled explicitly; (ii) it needs a formulation of the value of firms that involves adjustment costs of capital, or non-competitive practices, producing something more like a Lucas tree model of firms, for which values and returns fluctuate greatly; and (iii) it needs a financial sector. Such a treatment goes well beyond the present text and, moreover is challenging to put together, since it also needs confront the asset pricing puzzles discussed in Chapter 16. Below, we will therefore treat return processes as exogenous, when they are discussed.

21.2.2 Wage inequality

We now turn to earnings inequality, in particular that driven by differences in wages. The labor market is highly complex both in terms of how the overall market works and, given the market structure and prices, how individual workers fare relative to each other. We will focus mainly on general-equilibrium effects here and on markets with perfect competition. Of course, the labor market is not characterized by perfect competition and we will briefly comment on departures from it at the end of the section, but we still believe that the perfect-competition perspective gives a very useful benchmark and will, in many cases, give highly relevant practical insights.

Skill-biased technological change

The literature on wage inequality gained momentum as a result of the significant increase in the "skill premium" that started in the mid-1970s, as first documented by Katz and Murphy (1992). The general idea in this literature is that, given an aggregate production function $F_t(k, u, s)$ where u is unskilled and s skilled labor and the t subscript on F denotes a general form of technical change, one simply computes the relative wage of skilled and unskilled workers as the ratio of partial derivatives, $F_{s,t}/F_{u,t}$, evaluated at the aggregate quantities of labor and of capital. The literature has made different assumptions on the shape of F and the way in which technological change enters, leading to different interpretations of the data.

Katz and Murphy proposed an aggregate production function $F(k, G(A_u u, A_s s))$, i.e., a function where there is a sub-nesting G that only involves the two kinds of labor and where G has constant shape over time while technological change occurs through the As: skill-augmenting factors. The sub-nesting in particular means that the aggregate capital stock does not influence relative wages. With competitive input pricing, we obtain a skill premium

$$\frac{w_s}{w_u} = \frac{A_s G_s(A_u u, A_s s)}{A_u G_u(A_u u, A_s s)} = \frac{A_s}{A_u} \frac{G_s\left(1, \frac{A_s s}{A_u u}\right)}{G_u\left(1, \frac{A_s s}{A_u u}\right)}.$$

We see that both A_s/A_u—"skill-biased technological change"—and s/u—the relative quantities of the two inputs—matter. With G being a CES function, as Katz and Murphy further assume, one obtains

$$\log(w_s/w_u) = -\frac{1}{\rho}\log(s/u) + \frac{\rho - 1}{\rho}\log(A_s/A_u) + \text{ constant}.$$

As the relative supply of skilled workers increases, we see that the skill premium falls. We also see that skill-biased technological change makes the skill premium rise if and only if the substitution elasticity between the two skill types is above one. This is a standard result: when factor-augmenting technological change is biased toward one input factor, its relative productivity will fall if the two inputs have a sufficiently low degree of substitutability: due to the complementarity, the other factor becomes in higher "need". (With a Cobb-Douglas function, factor-augmenting technological change cannot affect the relative input prices as the direct productivity effect and the need effect cancel exactly.)

Katz and Murphy further write $A_s/A_u = (1+\gamma)^t$, i.e., interpreting there to be a constant rate of skill-biased technological change. With $\log(A_s/A_u) = t\log(1+\gamma)$, they then run the regression

$$\log(w_s/w_u) = -\beta_1 \log(s/u) + \beta_2 t + \text{ constant} + \epsilon.$$

Here, $\hat{\beta}_1$ allows us to back out an estimate of the substitution elasticity ρ and, based on this estimate, $\hat{\beta}_2$ can then be used to estimate of the rate of skill-biased technological change. To implement this procedure, one needs to define the notion of skill; Katz and Murphy use educational groups and add up to two labor inputs, one with college graduates and one with the remaining workers.[14] The supply of high-skilled workers increased quite fast during the period under study, pushing the high-skilled wages down. The regression estimate then implies a ρ around 1.4 and a counteracting demand force of around 10 percent per year, i.e., a very rapid rate of skill-biased technological change.[15]

Endogenous skill-biased technological change

The hypothesis of skill-biased technological change can be explored further, in part by trying to understand its determinants. One idea is that of *endogenous, directed technological change*, where the input-augmenting technology levels are derived as a function of changes in the environment. One such change is the relative supplies of labor themselves. Acemoglu (1998) proposes such a theory, along the lines of the endogenous growth literature discussed in Chapter 13 where purposeful R&D develops patents that are used to improve the productivity of skilled workers. This group having become larger thus constitutes the motivation for patent developers, who can then increase their profits by targeting that growing market. A simple version of this idea can be implemented as follows. Suppose the overall technology in society is given by

$$\max_{\{A_s,A_u\}} \left[(A_s s)^\sigma + (A_u u)^\sigma\right]^{1/\sigma} \text{ s.t. } \left[\lambda A_s^\phi + (1-\lambda) A_u^\phi\right]^{1/\phi} = 1,$$

where $\phi > 1$ to ensure an interior solution. The constraint describes the choice of the direction of technology: it builds in a trade-off between A_s and A_u, and this can be thought of as a reduced form of having to allocate a given amount of researchers to two different

[14] They use CPS data 1967–1987 and create education cells, each associated with an average wage within the cell. Then aggregation over cells uses these wages as weights, thus delivering a notion of efficiency units of labor input for skilled and unskilled labor, respectively.

[15] The value 1.4 was in line with the pre-existing literature of a "$\sqrt{2}$"-rule of thumb. However, Katz and Murphy also try other values for ρ and back out a skill-biased technology path year by year, yielding similar qualitative conclusions.

activities. Here, the elasticity of substitution between the two inputs for *given* input-saving technologies, ρ, equals $1/(1-\sigma)$. The first-order condition becomes

$$\frac{A_s}{A_u} = \frac{\lambda}{1-\lambda}\left(\frac{s}{u}\right)^{\sigma/(\phi-\sigma)}.$$

So if $\sigma \in (0,1)$ (more substitutability than Cobb-Douglas), this formulation delivers that a higher relative skill supply attracts more R&D. This means that a higher relative supply of skilled labor will, through endogenous directed technology, lead to a counteracting positive effect on the skill premium. Can this effect overturn the direct ("neoclassical") effect?

Substitute the expression for relative efficiencies into formula for relative wage (MPL_s/MPL_u) to obtain

$$\log\left(\frac{w_s}{w_u}\right) \propto \frac{\sigma - \phi\,(1-\sigma)}{\phi - \sigma}\log\left(\frac{s}{u}\right).$$

We see that the skill premium is *increasing* in $\frac{s}{u}$ as long as $\rho - 1 = \frac{\sigma}{1-\sigma} > \phi > 1$. I.e., with a large enough elasticity of substitution between skilled and unskilled labor (at least above 2), an increase in the relative supply of skilled labor, such as during a college attendance boom, can in fact increase the relative wage of college graduates.

Capital-skill complementarity

A related hypothesis is that skilled labor and unskilled labor play distinct roles in production and interact with capital in different ways. The hypothesis that goes back to Griliches (1969), who noted systematic differences in the average level of education and capital intensity across industries. Thus, the behavior of the skill premium over time could potentially be linked to an observable, namely, the behavior of capital, rather than be measured residually. Krusell, Ohanian, Ríos-Rull, and Violante (2000) pointed out that there had been especially fast investment-specific technological change (see Chapter 13) during the second half of the period studied by Katz and Murphy, precisely when the wage skill premium started rising.[16]

These ideas can be formalized by adopting a slightly different production function, nesting capital asymmetrically with the two types of labor: conceptually, $F(u, G(k,s))$ allows us to use a higher substitutability in the G nest than in the F nest. Thus, rapid growth in k can make the relative wage of skilled labor go up without technology factors playing a role.

To understand the logic, consider the simple example $F(k,u,s) = u + s^{\nu}k^{1-\nu}$. Clearly, higher capital raises the marginal product of skilled labor, whereas the unskilled marginal product is unaffected (and equal to 1). Generalizing to both F and G being of the CES kind, one can show that if the elasticity in F is greater than that in G, then the skill premium rises as k goes up. The account provided by Krusell et al. (2000) does suggest capital-skill complementarity has played an important role.[17] By this account, where the growth of equipment capital raises the productivity of skilled labor more than that of unskilled labor, we are closer to direct measurement of the source behind skill-biased technological change.

[16] Their study focused on equipment capital, as opposed to structures. For simplicity, we use a general notion of capital here.

[17] Ohanian et al. (2023), with recent data, estimate the higher elasticity to be slightly below 2 and the lower one slightly above 0.5.

A different nesting, $F(s, G(k, u))$, allows a similar interpretation but some different qualitative properties. In particular, a rise in the capital stock now allows the real wage of unskilled workers, w_u, to fall (e.g., consider $s^\nu(k+u)^{1-\nu}$). This feature, which cannot be obtained under the previous nesting, is in line with the stagnant wages of parts of the population over several decades beginning in the 1970s.

Finally, capital has also been argued to be especially complementary with human capital in times of rapid technological progress, as in Greenwood and Yorokoglu (1997). The idea here is that human capital increases a worker's ability to adapt to new circumstances.

Human capital accumulation

The previous sections emphasize skill differences as important determinants of wages, in particular as measured by formal education. More generally, we conceptualize "human capital" as a key determinant of wages. What is human capital, and how is it accumulated?

There is a vast literature covering many aspects of human capital relevant for labor markets. First, human capital can be a thought of as a general skill, useful in many tasks, occupations, and industries in generating more output per unit of time. Such a skill can be accumulated during education as well as while working. But skills can also be quite specific and not easily transferable across occupations or jobs. More generally, human capital is multi-dimensional, in which case it is natural that workers sort across jobs and tasks. The Roy model (Roy, 1951) expresses this clearly: each individual has a skill vector describing the amounts of different types of skills, and different jobs give different returns to the different skill types; the model then describes, given a distribution of individuals across skill vectors, how individuals sort into jobs to maximize their income given their skill combinations. Furthermore, even when skills are one-dimensional there can be a non-trivial pattern by which individuals match to different job features; for example, the skill can be complementary with capital, whereby markets would push toward high-skilled individuals matching with firms that have advanced capital equipment. This hypothesis is closely related to the capital-skill complementarity hypothesis earlier in this section.[18]

There is also a rich set of models of the accumulation of human capital. These models are typically framed in a partial-equilibrium setting but are often imported into macroeconomic general-equilibrium settings. A basic model of education was provided in Mincer (1974), where the key choice is how many years of schooling to obtain. Under some conditions, the framework delivers the well-known Mincer equation. This equation characterizes the log wage of an individual as linear in the number of years of schooling and in the years of work experience; when applied econometrically, it delivers a "return to schooling" (the coefficient on the years of schooling) of around 0.1 in a large number of contexts. Ben-Porath (1967) gives us a basic, more general model of human capital accumulation, where one's time can be divided into working or human capital accumulation (education or training). The framework naturally gives rise to a specialization on human capital accumulation early on and a specialization on working only later in life, when investments in human capital no longer pay off given the short remaining working life; midlife, there is an interior solution with both work and training.

[18] High-skilled individuals can also be complementary with other high-skilled individuals, such as in the O-ring theory of Kremer (1993).

Both the models of human capital accumulation just noted are particularly simple in that they assume a fixed amount of time available (for either work or study), along with perfect credit markets. This means that the objective function can be phrased as present-value lifetime income: there is a separation between this problem and the smoothing of consumption over time. If credit markets are not perfect, the human capital accumulation decision, as discussed briefly in Section 21.3.1 below, becomes intertwined with the consumption decision. Hence, the individual's wealth level matters, and a worker can then make decisions that appear suboptimal in a present-value sense—e.g., they fail to educate themselves or they take jobs involving no training—because they lack liquidity and need cash for consumption purposes; even low-skill traps are conceivable. See Griffy (2021) for a recent study.

Task-based models

The approach based on aggregate production functions skips the many details of actual production processes and how different individuals sort into different occupations and tasks to be performed across the different sectors of the economy. A recent line of research offers an alternative abstraction: a production process is defined by a set of tasks to be performed, and the tasks are all complementary (see, e.g., Acemoglu and Autor, 2011). Then different inputs, such as labor of different skill types and capital, have different abilities to produce different tasks. For each task, the different inputs are substitutable; in the simplest case, they are perfect substitutes, with coefficients that differ by task and input. In a competitive environment, the different inputs are then allocated optimally to different tasks. A full model description and solution is beyond the scope here; suffice it to say that there are specific assumptions on the task/input coefficients such that one can derive a closed-form expression for overall production as a function of the amount of inputs (capital and the different labor inputs). Thus, an aggregate production function is derived endogenously and its properties depend on the underlying task/input assumptions.

The task model is appealing in that it can give a concrete expression for concepts such as "automation". A prime example is the "hollowing out" of the employment distribution in the U.S. during the last century's last decades, when many middle-manager tasks, according to a common account, were automated.[19] That is, specific worker skills were made obsolete. A topic of current interests is the adoption of robots: a large number of tasks currently performed by labor may instead performed by capital (say, because capital has become more abundant, or because the task/input features changed). Again, it is a worker's type of skill that determines whether it is easy to replace by machines or not. The task model also allows one to analyze *outsourcing*, i.e., the idea that some tasks in a production process is performed by workers abroad, where wages may be lower. When the model is applied empirically, one needs to identify tasks and usually, the O*NET data—Occupational Information Network (a U.S. government initiative), linking skills to job requirements—is then used; the task descriptions "abstract", "routine", and "manual" are then mapped into how easily they can be performed by different labor (or capital) inputs.

[19] See, e.g., Acemoglu and Autor (2011).

Labor markets in practice

The above discussion is based on aggregate production functions and the assumption that wages equal marginal products. In practice, labor markets of course have many features that influence wages through other mechanisms. We now briefly mention examples of such mechanisms.

Search frictions In Chapter 20, we saw that wages for identical workers can differ due to search frictions. It is hard to assess the importance of this channel as the notion of "identical workers" receiving different wages is hard to implement in practice. Measures based on "wage residuals", i.e., wage regressions using all available observables still deliver large residual dispersion, but personal characteristics such as "diligence" or "ability to cooperate" may still differ greatly within the set of individuals with the same observable characteristics. Significant wage dispersion due to search frictions alone also raise the issue of why some employers pay so much more than others.

Compensating differentials Another important feature of labor markets is that workers value other aspects of jobs than just the earnings: they take into account the *amenities* of different jobs. Today, a large number of job amenities are in principle available for many employees (just google "job amenity" and you will see a large number of examples!), who then may accept jobs at lower wages than otherwise; the possibility to work at home is one that is very popular today but only seems to have become common after the Covid pandemic. Similarly, some jobs have negative amenities like lack of job safety or highly irregular work hours, but what appears like a plus to some workers may be a minus to others, so there is also selection to take into account.[20] From a welfare perspective, we note that wage dispersion due to the amenity heterogeneity of jobs is a natural, and arguably desirable, feature: it would seem undesirable to have equal wages at two jobs that are identical but differ markedly in the attractiveness of their amenities.

Monopsony and union power Another currently very active research area is to assess the role of monopsony power in labor markets: firms' abilities to lower wages below marginal products because, for one reason or another, it is costly for workers to change jobs. Like the presence of job amenities, monopsony power is multi-faceted and hard to measure; the key from the perspective of understanding wage inequality is how different firms benefit from different degrees of monopsony power and the observation of rising firm concentration over the last several decades has directed the attention of researchers to this possibility.

Another departure from wages equaling marginal products is made possible by workers having a degree of market power vis-à-vis firms. Unions are a central component of labor markets in Europe, while much less so in the U.S. today; in the U.S., union membership peaked in the 1950s at a little over 30% of workers being unionized, whereas the number today is below 10%. Moreover, in Europe also many non-unionized workers are covered by so-called collective agreements, i.e., contracts negotiated between unions and employer

[20] As increasing amounts of micro data on job descriptions has become available recently, empirical research on amenities is currently very active.

federations stipulating a range of features of the labor contract, including wage floors and amenities. Changes in the degree of influence that unions exert on labor markets are broadly viewed to be important determinants of wages and wage inequality, though hard evidence in the form of controlled experiments is only scant. Before the search and matching theory became the dominant framework for analyzing labor markets and unemployment, theories based on union wage determination were in focus. Today, we see a certain resurgence of such theory, at least partly because we are now able to access more data on wage contracts and union membership.

21.2.3 Wealth inequality

In this section we will look at a number of determinants of wealth inequality. We will focus on the long run, i.e., we will use a number of models evaluated at their steady states as a way of relating to the data described in the earlier sections.[21] This seems reasonable as the wealth distribution has certain properties—it is highly skewed and it is significantly more dispersed than the earnings distribution—that hold true over time and across countries. There has been significant research in macroeconomics over the last decades aiming at accounting for this feature of the data; as we shall see, it is a challenging problem.

Recall from above that the relative shares of capital and labor income can be analyzed straightforwardly using our aggregate production function. Here, however, the focus is on differences in capital holdings between households. In the most commonly used model of wealth inequality today—the heterogeneous-agent model—consumers are subject to idiosyncratic, uninsurable shocks, and they accumulate and decumulate their asset position in part to self-insure against shocks. For a thorough understanding of this class of models, however, it is important to first understand how a model without shocks works. We therefore first look at deterministic models. We will begin with the simplest possible model and then gradually introduce more elements.

Deterministic models

We begin with the frictionless dynastic model, and we will see that this model has very particular long-run predictions for asset inequality: any long-run wealth distribution is possible, and which one will occur is entirely dependent on the time-zero wealth distribution. We then discuss various extensions that can potentially break this result.

The benchmark frictionless model We assume that all consumers have the same preferences but begin with different wealth levels. They also have different earnings in the form of different efficiency units of labor; labor supply is assumed to be inelastically supplied (the measure of consumers is 1 and each consumer supplies 1 labor unit). There is no uncertainty. In a steady state, we thus have consumer i maximizing

$$\sum_{t=0}^{\infty} \beta^t u(c_{i,t})$$

[21] How wealth inequality varies over time, out of steady state and in response to exogenous aggregate shocks, is not addressed in this section but are briefly commented on below.

subject to

$$c_{i,t} + a_{i,t+1} = (1 + r - \delta)a_{i,t} + \epsilon_i w$$

for all t.[22] Here, ϵ measures the number of efficiency units of labor and we assume that $\sum_i \epsilon_i = 1$. Hence, total labor input also equals 1. We assume a standard neoclassical model, for simplicity without growth, with a perfectly competitive firm sector. Aggregate capital at t is given by $k_t = \sum_i a_{i,t}$ at all times given that we assume a closed economy and as capital is the only asset in positive net supply.

The steady-state level of capital is found by evaluating the consumers' Euler equations. They all face the same interest rate and they have the same discount factor so, given $c_{i,t} = c_{i,t+1}$ for all i, the Euler equation is identical and satisfied for all consumers when $\beta(1+r-\delta) = 1$. The firm's first-order conditions for profit maximization deliver r and w as a function of aggregate capital, which allows us to pin down the steady-state level of capital, k. This, in turn determines w as well. What remains is to find the distribution of consumption and asset levels for consumers. The only remaining equations to use are consumers' budget equations, which read $c_i = (r - \delta)a_i + \epsilon_i w$ for each i. We note that for each agent there is one equation but two unknowns, c_i and a_i; hence any combination is possible, so long as $c_i \geq 0$. We thus have steady-state *indeterminacy* in the wealth and consumption distributions. In particular, labor earnings are not connected to wealth or consumption.

To understand why we obtain indeterminacy, recall that we demonstrated at various points earlier in the text, that we have a standard permanent-income setting: the consumer maintains the initial asset level forever, thus consuming earnings and the net interest income off of the wealth. Higher initial wealth thus simply produces higher consumption, with a marginal propensity $r - \delta$.

We can imagine that consumers also receive shocks to earnings. Below we will look at the effects of such shocks when they are not fully insurable. But assume now that they are fully insurable and idiosyncratic, without any associated aggregate risk. Then steady-state indeterminacy would obtain again: whatever it is, the initial wealth distribution stays constant over time, provided it sums to steady-state k, and the consumption distribution follows mechanically. Budget constraints are more complicated due to the presence of insurance instruments but consumption remains constant and insulated from shocks.[23]

Departures from a dynastic setting Consider now an overlapping-generations model, where agents live for a finite number of periods and do not give bequests (because they do not value their offspring). We know from earlier in the book that the steady-state interest rate is nontrivially pinned down; its value depends on the life-time earnings profiles of agents and their implied savings needs given their desire to smooth consumption and, in the presence of capital accumulation and production, in conjunction with the properties of the production function. All consumers save zero in the final period and if they all start with zero wealth (which is natural given the absence of bequests), the relative wealth holdings during their lifetimes will depend on the timing and size of their earnings; agents with high earnings that occur early save a lot and become the wealthiest, but if these agents's earnings occur late

[22] We also assume the standard no-Ponzi-game restriction.

[23] If the initial capital stock is not at steady state, the initial asset distribution still determines the long-run distribution but it evolves non-trivially over time. A full characterization is available in Chatterjee (1994).

then they will borrow and become the wealth-poorest. In sum, the wealth distribution will be determinate and will reflect the earning profiles of agents during their lifetimes.

Suppose now that, in the same overlapping-generations model, we add altruism: each agent values their offspring, as modeled by an additional term in utility multiplied by $\beta < 1$, thus assigning a smaller value for children. With two-period-lived agents and life utility $U(c_{y,t}, c_{o,t+1})$ for an agent who is young at t, we obtain an indirect utility function

$$V(a_t) = \max_{c_{y,t}, c_{o,t+1}, a_{t+1}} U(c_{y,t}, c_{o,t+1}) + \beta V(a_{t+1}) \quad \text{subject to}$$

$$c_{y,t} + \frac{c_{o,t+1}}{1+r-\delta} + a_{t+1} = a_t(1+r-\delta) + w(\epsilon_y + \frac{1}{1+r-\delta}\epsilon_o),$$

where we find it convenient to define the bequest from generation t to generation $t+1$ as a_{t+1}, an amount given at t already and received at $t+1$, and where we have assumed steady state and that all generations have the same earnings profile. This problem can be rephrased as

$$V(a_t) = \max_{R_{y,t}, a_{t+1}} u(R_{y,t}) + \beta V(a_{t+1}) \quad \text{subject to}$$

$$R_{y,t} + a_{t+1} = a_t(1+r-\delta) + w\hat{\epsilon},$$

where $u(R) \equiv \max_{c_y, c_o} U(c_y, c_o)$ subject to $c_y + c_o/(1+r-\delta) = R$ and $\hat{\epsilon} = (\epsilon_y + \frac{1}{1+r-\delta}\epsilon_o)$. This is now an entirely standard permanent-income model, and hence any conclusions we obtained for the dynastic household apply here as well. In particular, the steady-state wealth distribution is indeterminate. We thus see that the assumption of altruism plays a crucial role.

There are also other formulations that involve bequests. One is the "warm-glow" setting, where an agent receives utility from giving bequests—a joy of giving—according to an exogenous function; let us denote it by $\hat{V}$. Notice that this is in sharp contrast with V above, which is endogenous and would change, say, if the returns to saving or policy changed. V would also change in response to differences in earnings across cohorts; earnings-rich parents might want to give more bequests to earnings-poor children, for example. Thus, the warm glow formulation is non-standard in that it simply enters an action—an amount of saving—into the utility function as a primitive, unlike a consumption good where prices influence demand.[24]

Discount factor heterogeneity

Still for a frictionless framework, suppose consumer i has a discount factor β_i, with $\beta_i \neq \beta_{i'}$ for some (i, i'). Then the long-run wealth distribution is determinate and very special. First

[24] As a non-generic outcome, it is possible that $\hat{V}$ and V coincide for two economies that are otherwise identical. However, a change in policy or, say, earnings structures, would change V but not $\hat{V}$, and they would no longer coincide. In particular, the warm-glow model does not map into Arrow-Debreu, and hence standard welfare theorems do not apply; see Chapter 6 for an example with dynamic inefficiency. Consequently, what is assumed about $\hat{V}$ becomes critical in terms of the long-run implications for wealth inequality. In particular, what is important is the relative curvature of $\hat{V}$ relative to the parent's utility function of consumption. Taking prices as given, a relatively less curved $\hat{V}$ translates into higher future wealth inequality given current wealth inequality; however, in general equilibrium prices may counteract this initial effect.

of all, the consumer (i^*) with the highest discount factor has constant consumption and the capital stock is determined by $\beta_{i^*}(1+r-\delta)=1$, while all other consumers' consumption levels are converging to zero, i.e., they do not have constant consumption. Thus, all the other consumers have their wealth levels given by $0=(r-\delta)a_i+\epsilon_i w$, i.e., a negative level such that consumption equals zero.

Intuitively, more patient consumers save more and, in the limit, though convergence will be slow to the extent that the differences between discount factors are small, the most patient consumer consumes the economy's entire output.[25] We also note that the extreme long-run inequality is a result of choice. Thus, it is hard to argue that it is "unfair": those who consume very little later on have simply chosen to consume more early. From this perspective, wealth inequality per se should not be seen as undesirable.

An assumption of exogenous and permanently different discount factors—or permanent differences in other preference parameters—is arguably not an appropriate assumption for a dynastic model. A reasonable alternative is to allow randomness that can capture how preferences are rather stable during a person's lifetime but less so between parents and children.

Distortions, credit-market restrictions, and imperfect asset markets

We now briefly look at deterministic environments with distortions or restrictions on choice. Beginning with distortions, suppose there is a tax on capital income (net of costs of capital) that is progressive: gross capital income is $a\,(1+(r-\delta)(1-\tau(a)))$, where $\tau(a)$ is the tax rate schedule as a function of the amount saved. So suppose $\tau'(a)$ is positive and strictly increasing. Then the Euler equation, which now involves $\tau(a)$ and $\tau'(a)$, cannot be met for a constant (steady-state) consumption path at the same time for all agents, to the extent they have different levels of a.[26] In this sense, the situation is quite like that under discount-rate heterogeneity. However, the outcome here is the reverse, i.e., a wealth distribution that converges to full equality over time: in steady state, $a_i=a_{i'}$ for all i and i'. Intuitively, progressive taxes on capital income simply slowly eat away the wealth of savers, and the more so the higher the level of savings. Proportional taxes on capital income also lower saving, but not differentially across saving levels.

The case of a progressive tax on capital income, or one on wealth, is likely important in practice for understanding the long-run evolution of inequality. Hubmer et al. (2018) in particular argue that Reagan's tax cuts and decreases in the degree of tax progressivity, which subsequently have not been reversed, constitute a key factor behind the slow further build-up in wealth concentration.

Turning to credit-market restrictions, we shall see that these can, when they restrict borrowing for consumption, lower asset inequality. Consider a setting with two representative dynastic consumers, A and B, equal in numbers, where consumer A has endowment $\bar{y}$ in even periods and $\underline{y}<\bar{y}$ in odd periods; for consumer B the situation is the reverse (high

[25] It is convenient for these reasons to assume that there is a large number of consumers with each discount factor; otherwise the assumption of price-taking is not appropriate. Within each discount factor group, then, the wealth distribution is still indeterminate.

[26] The gross return on saving now equals $1+(r-\delta)(1-\tau(a)-\tau'(a)/a)$. Thus, in steady state this expression times β must equal 1.

endowment in odd periods, low endowment in even periods). Each agent's budget constraint is $c_t + q_t a_{t+1} = y_t + a_t$. Total consumption equals total endowments in each period; there is no production. With a standard utility function $\sum_{t=0}^{\infty} \beta^t u(c_t)$ for both agents, where u is strictly concave, let us find the equilibrium interest rate and the resulting pattern of borrowing and lending. First, then when there are no restrictions on borrowing, there will be full consumption smoothing and $q = \beta$ (the gross interest rate will equal $1/\beta$). This is achieved by the currently endowment-rich agent lending to the endowment-poor agent each period, so that each consumer is a borrower one period and a lender the next period.[27] This outcome can be prevented by the existence of borrowing constraints. Suppose $a_{t+1} \geq \underline{a}$ at all times, with $\underline{a} < 0$ close enough to zero that the constraint binds at all times. Then in equilibrium, consumers are not able to fully smooth, because they are not allowed to let their assets move freely. That is, asset inequality is a sign of an imperfect asset market, and consumers are worse off. The interest rate from periods 1 and on will be determined by

$$qu'(\bar{y} + \underline{a}(1+q)) = \beta u'(\underline{y} - \underline{a}(1+q)),$$

which is one equation and one unknown, given that $\underline{a}$ is exogenous.[28]

In the data, as we have seen above, there is a significant fraction of households with negative financial asset holdings. This is in part due to consumption loans, in a manner similar to that just described, but it can also be due to borrowing to fund investment (e.g., in housing, via mortgage loans). In the case of investment loans, restrictions would not affect asset inequality much; they would simply distort the portfolio choice, which in turn can have real consequences, such as for how and where to live.

Finally, a potentially very important determinant of wealth inequality is in place to the extent that different consumers simply obtain different returns on their wealth in asset markets. This phenomenon is challenging to fully understand without market imperfections, asymmetric information, commitment problems, or behavioral components. At this stage, the macroeconomic literature has only begun to explore this channel, as there is no quantitative, off-the-shelf model of return heterogeneity. At the same time, with increased access to individual data on asset holdings and asset returns, we see that return heterogeneity is quantitatively significant. Some of this heterogeneity is due to different households holding different kinds of portfolios (e.g., housing, liquid savings in the form of bank deposits, publicly traded stock, private equity, or cryptocurrency) but some is due to heterogeneity within asset class; for example, investments in individual stock, as opposed to a market index fund, is commonplace and, clearly, contribute to wealth inequality. Clearly, those who invested early in cryptocurrency made a rapid climb upward the wealth distribution and, more generally, rapid movements in a household's relative wealth position are often explained through risky asset investments.

The simplest model of return differences in terms of the model above is to assume, ad hoc, that different consumers receive permanently different deterministic returns $r_i - \delta$ on saving, where i denotes a specific consumer. One can close the model by assuming that

[27] The exact amount of borrowing depends on the present-value wealth for the two consumers, which in turn depends on the time-zero asset position. If $a_0 = 0$ for both agents, then agent A is richer in present-value terms since they receive the high endowment first, at time 0, and will therefore permanently have a somewhat higher consumption that agent B.

[28] In the very first period, the interest rate is different if the initial asset position is $a_0 = 0$ for both agent.

$\sum_i r_i a_i = r \sum_i a$, where r is the marginal product of capital in production, evaluated at $k = \sum_i a_i$.[29] Simple inspection of the Euler equations of consumers tells us that the agent with the highest r_i must hold the entire capital stock plus an amount of lending to the remaining consumers, who have zero consumption and negative asset holdings.

Another, more structural description of return heterogeneity is an environment where an "entrepreneur" has access to a high-returning investment project but credit markets are restricted, along the lines of Chapter 19. If markets worked perfectly—if the entrepreneur could obtain funds at a frictionless credit market—then others could share in obtaining the high return and hence the project would just be part of the overall production possibility set of the economy; there would be no sense in which the entrepreneur could obtain a higher return on saving than anyone else. So imagine instead that such capitalization of the investment project were not possible but instead the entrepreneur had to self-finance, provided a minimum investment amount is attained from the investor's own saving. Such a setting, which was analyzed by Quadrini (2000), generates high wealth accumulation among those with enough money and opportunities to invest. Which projects are funded by banks, by private equity, by bond finance, and through public stock exchanges, varies greatly across time and countries; the point here is that the efficiency and specific nature of these markets matters for wealth inequality.

Models with idiosyncratic, uninsurable shocks

We now turn to the class of heterogeneous-agent models of wealth inequality that also form a new core of much of modern macroeconomic analysis. We begin with the simplest such framework and then discuss extensions.

Idiosyncratic earnings shocks and precautionary saving If earnings shocks are idiosyncratic but not insurable, except for the possibility of saving in a riskfree asset, then we have the settings analyzed by Aiyagari (1994) and Huggett (1993b) and described in Chapter 11. The first paper in this literature, however, was Imrohoroğlu (1989): she demonstrated how business cycles affect the welfare costs of business cycles through the effects on consumption inequality.[30]

Relative to the results above, in the present section we note that the presence of uninsurable, idiosyncratic shocks makes the steady-state distribution of wealth determinate. This is most easily seen using the Huggett version of the model, where we recall that the budget constraint reads

$$c_t + q_t a_{t+1} = \epsilon_t + a_t,$$

where ϵ is idiosyncratic and random and there is also a borrowing constraint: $a_{t+1} \geq \underline{a}$. There is no production or aggregate storage, so assets always sum to zero in the population. Given the possibility of borrowing, some agents will want to borrow and others lend, so long

[29] For the summing up of income to work out, the r_is need to be endogenously connected to the asset distribution, as the sum of capital incomes needs to match $r \sum_i a_i$; clearly, one needs to understand just how the return differences materialize in order for a complete, general-equilibrium understanding of this phenomenon.

[30] Imrohoroğlu (1989) assumes that the real interest rate is exogenous and constant; Aiyagari (1994) and Huggett (1993b) derive the interest rate endogenously, but do not have aggregate shocks.

as ϵ is mean-reverting, such as an AR(1) process with autocorrelation strictly less than 1. We saw the model applied to a special case in Chapter 17: the case of $\underline{a} = 0$, where borrowing is not allowed and the equilibrium is autarky, with a real interest rate determined in closed form based on the Euler equation for the non-constrained agent. This example makes the point that very unevenly spread out wealth can reflect something positive: households are able to share risk through borrowing and lending. The extreme opposite case—the autarky outcome, which has perfect equality in asset holdings—allows no consumption smoothing at all. The most extreme asset inequality is obtained when $\underline{a}$ is set at the natural borrowing limit: the most generous level such that debt can always be paid back. It also shows that the level of the interest rate is a key variable in the macroeconomic determination of asset inequality.

In an Aiyagari model, there is also aggregate saving, with returns influenced by a neoclassical production function. From the perspective of the individual's need to insure, this offers an additional possibility, as explained in Chapter 11. As a consequence, with more risk to be insured, or higher risk aversion, the steady-state capital stock will rise, reflecting the increased need to save.

Quantitatively, a simple AR(1) process for earnings generates wealth inequality that is more dispersed than earnings inequality, but not by a sufficient amount to match high wealth concentration at the top as observed in the data. The reason for this is that the richest do not value insurance so much—by virtue of being rich, as their accumulated saving provide a very good buffer against bad earnings shocks. In addition, as we saw in Chapters 11 and 17, in this class of economies, the riskfree rate is below the discount rate, which means that well-insured agents will *decumulate* wealth. Thus, the right tail of the wealth distribution becomes limited: the precautionary need dies off at higher wealth levels and the low return on savings dominates.

If, on the other hand and as in Castañeda, Díaz-Giménez, and Ríos-Rull (2003), the earnings process has a combination of (i) very large upside shocks while (ii) still allowing a non-trivial risk to fall far down from the top of the earnings distribution, then the model generates highly a skewed wealth distribution. Very large earnings shocks and consumption smoothing together makes it possible to accumulate large amounts of wealth possible, and the decumulation at high levels is hampered by the precautionary motive still being active, since earnings can drop precipitously.

Idiosyncratic shocks to discount factors or asset returns We saw above that permanent differences in discount factors generate a determinate, and extreme, wealth distribution. Now consider random movements in discount factors, say, at the frequency of cohorts; e.g., as the dynasty wealth is passed on from one cohort to the next, the new dynasty head may be more, or less, patient than the previous one, and in a random manner.[31] Suppose these shocks are idiosyncratic and independent across dynasties, possibly with some persistence, but drawn from the same distribution; then the steady-state wealth distribution will be more

[31] Formally, in a model with constant (geometric) discounting, the discount factor applied to utils at t in terms of utils at 0 is β^t; here, they would be $\prod_{s=0}^{t-1} \beta_s$, where β_t is random, e.g., a first-order Markov process. This kind of discounting is time-consistent, i.e., the consumer would not want to change a contingent plan made in advance when the future arrives.

dispersed than in the case of common discount factors.

Next, suppose individuals receive shocks to the returns of their portfolios, again in an idiosyncratic manner but drawn from the same distribution; these shocks could also feature some persistence. Then again the wealth distribution would spread out and the higher is the variance and persistence of the shocks, the more the steady-state wealth distribution will spread out.

Let us now take a very brief detour from the structural models in focus in this section: suppose we simply have a framework of "random growth", as described in Kesten (1973):

$$a_{t+1} = s_t a_t + \epsilon_t,$$

where s_t and ϵ_t are i.i.d. Then under some conditions on the primitives of this process, it turns out that a_t converges in probability to a random variable A that satisfies $\lim_{a\to\infty} Prob(A > a) \propto a^{-\zeta}$, i.e., the right tail of the stationary distribution has a Pareto shape.[32] This result is remarkable in that the processes for s and ϵ are unspecified, and yet the limit distribution is of a specific shape, namely Pareto, which is also a very good approximation to the right-most tail of wealth, as well as earnings, in the data.

It is possible to connect the random growth model to our structural model, allowing for idiosyncratic shocks to earnings, discount factors, and returns. It turns out—for details, see Hubmer et al. (2018)—that if one assumes CRRA utility, then for large asset positions, the random-growth formula is a very good approximation to optimal behavior of saving.[33] In particular, the randomness in s has two components, one deriving from the discount factor shock while the other is a return shock: intuitively, the money saved is multiplied by a gross return, which is random, to which a marginal saving rate, also random due to discount-factor heterogeneity, is then applied. The ϵ, then, contains the earnings shock. Note that the Pareto shape will not apply unless s is random, so earnings shocks per se do not generate Pareto tails, unless they are Pareto distributed themselves.[34]

Finally, notice that the two mechanisms behind a right Pareto tail—random discount factors and random returns—are quite different from a welfare perspective. In the former case, the wealth distribution reflects conscious choice: poor dynasties are those that, on average at least, had low discount factors earlier on. Thus, when born into a dynasty with low wealth, high current consumption is not an option but high consumption was probably what occurred earlier in time for this dynasty. Moreover, if offered insurance, a dynasty would not necessarily want to hedge a low (or high) future discount factor, as preferences throughout have been assumed to be time-consistent.

In contrast, random returns fundamentally lead to undesired wealth dispersion: those with high wealth are wealthy because they were lucky in the asset market. If given the opportunity to insure, the consumer would highly value such insurance. Thus, a currently poor dynasty is highly likely poor because the ancestors were unlucky in the returns on their savings.

[32]The result follows if there exists a $\zeta > 0$ with $\mathbb{E}[s^\zeta] = 1$, with $\mathbb{E}[\epsilon^\zeta] < \infty$. For a nice exposition, see Gabaix (2009).

[33]The linearity applies quite well also lower down in the distribution, except for the lowest levels of wealth where the asset evolution is noticeably non-linear with a slight convexity.

[34]In this case, we obtain a Pareto tail for assets equaling that for earnings.

Hours worked In the discussion above, earnings were taken to be exogenous. If individuals can choose how much to work, then this margin can affect wealth inequality. First, labor supply can be used as additional insurance vehicle in response to shocks. For some shocks, such as those to wages, then a positive shock can of course also lead to higher hours worked, in this case enhancing earnings variability. The interaction of hours worked and wealth accumulation also allows this class of models to address the data on the distribution of hours worked. While there is considerable variation in hours worked across individuals, it is difficult to understand the determinants of this variation since the correlation of hours with observable worker characteristics tends to be low. For example, the correlation between hours and wages or financial wealth is curiously close to zero, despite a common perception that productive or rich people on average work harder. A model with wage shocks and labor supply featuring strong income effects, in line with the discussion in Chapter 12, can make sense of a zero, or weak, correlation: income effects make the wealthier want to work less, but intertemporal substitution of labor supply at the same time generates high hours worked by those with high current wages and who also save in order to smooth consumption. In this kind of environment, tax policy distorting labor decisions, will also influence the observed correlations. For details, see Domeij and Floden (2006) and Pijoan-Mas (2006).

Why are so many so poor? In the models above, the focus was mostly on mechanisms through which the right tail of the wealth distribution becomes thick, as it is in the data. A different, but no less important, question is to understand the wealth formation, or lack thereof, of the very poorest. As we saw earlier in the chapter, the 40% poorest only hold 1% of total financial wealth, and in addition there is poor coverage of people in this part of the distribution.

A benchmark Aiyagari model predicts many too few individuals in the left tail of the wealth distribution, simply because they save their way out given that it is painful to have very low consumption. A key missing element is social security support—provided either by government in the form of transfers, food stamps, and other free goods and services, or by family/friends—that effectively constitutes a consumption floor, hence making it much less costly not to have wealth.

Many low earners may also lack effective means of saving, if they have needs to conceal wealth from others due to informal sharing arrangements within multi-person households or social networks. Yet others, who have defaulted on loans but have been unable to file for bankruptcy, may be able to save but their savings are then typically garnishable by creditors.

Clearly, many individuals may also suffer from mental conditions, addiction problems, or simply elements of irrationality not captured by the standard utility functions used here. Another element missing in our models is how crime, both as an activity and through incarceration, shapes the earnings and wealth of those involved. The very poorest are generally understudied in economics, including in macroeconomic analyses.

Quantitative analysis We now briefly illustrate the above points by presenting results from steady states given a set of extensions of the Aiyagari model. First off, a standard Aiyagari model, calibrated to PSID data using an AR(1) income shock, as in Aiyagari (1994), delivers an income Gini coefficient of 0.37 and a wealth Gini of 0.67, with a steady-

state interest rate of 2 percent and a ratio of capital to annual GDP of 3. Thus, the model does not generate nearly as much wealth inequality as we see in the data (recall that it is significantly above 0.8); the percentage of wealth held by the 1 percent richest is 9%, as opposed to nearly 40% in the data. The average MPC is 0.17, but the bottom 10% of the distribution has an MPC of 0.78. Let us now look at a number of extensions.

A superstar earnings process: assume earnings are in line with Castañeda, Díaz-Giménez, and Ríos-Rull (2003), such that the Gini coefficient for wealth equals 0.8; this involves an income Gini of 0.67, with the same interest rate and k/y ratio as in the standard model. The share of wealth held by the 1 percent richest is now over 0.4, i.e., it even overshoots. This model, however, while delivering highly dispersed wealth, does not change the MPC distribution in the population more than very marginally.

Stochastic discount factors: assume that β takes on 3 possible values randomly. This process is persistent, with expected duration of any given value of β of around 75 years, and has 10 percent of the population at each of the extreme values and 80 percent in the middle. The three β values are then chosen to match a wealth Gini of 0.8. This model generates more moderate wealth concentration at the top—the 1 percent richest have only 19% of all wealth—but depresses the real interest rate to around 0.5 percent (and a k/y slightly above 3): this rate is highly influenced by the most patient, who now have a high β. Here the MPC distribution is moved up significantly, to an average of 0.29; the poorest have much higher MPCs, while the richest still have very low MPCs.

Random returns to saving: assume that the returns are iid with a standard deviation chosen to match the 0.8 wealth Gini. Now the (average) real interest rate is 3 percent, with a k/y ratio a little below 3. The richest 1 percent hold a fraction right in between those of the previous two model versions (so around 0.3) but the MPC distribution is again back at roughly the same level as for the basic Aiyagari model.

Comparing all these models we see that different theories of wealth inequality imply very different MPC distributions. There is, as of yet, no firm consensus as to which mix of assumptions is most appropriate: the research is very much ongoing.[35]

A note on welfare comparisons in heterogeneous-agent models

We end this section by briefly discussing a conceptual challenge—one that is as natural as it is important—when welfare comparisons are made in economies that are dynamic and inhabited by a heterogeneous population.

Transition First, we already know from Chapter 6 that in a dynamic representative-agent model, comparing steady-state welfare (say, resulting from two alternative policies) is not sufficient: the transition path needs to be taken into account. For example, let the steady-state capital stock in a frictionless benchmark be k^*. Then a small, time-independent subsidy to capital income would take us to a new steady state in which capital, and hence welfare, are higher.[36] But taking the transition path into account, the welfare of the representative agent, at time 0 as of the introduction of the policy, must be lower, since the initial economy

[35] See Ozkan, Hubmer, Salgado, and Halvorsen (2023) for recent advances.

[36] Recall that in the standard dynastic model, steady-state welfare is not maximized: due to discounting, it involves slightly lower capital than that maximizing steady-state consumption.

is efficient. Intuitively in this case, the higher capital accumulation requires consumption to be lower during the initial phase of the transition, outweighing the benefits from the higher long-run consumption due to higher capital.

With heterogeneous agents, it is of course in addition important to take into account how different individuals are affected by a counterfactual experiment. In the benchmark model we will discuss below, consumers are also hit by various shocks that are not fully insurable; hence, they also move around in the distribution over time. This poses another conceptual challenge. Say that we are again interested in the effects on welfare of adopting a subsidy to saving, and say that the initial position is one of a steady state. Then the correct welfare comparison is arrived at by (i) solving for a full transition path given the new policy and (ii) comparing present-value utility for all agents, as of time 0 (taking transition into account). This may result in some agents gaining, and others losing, from adopting the subsidy. Such a result could be reported to a policymaker wanting to evaluate the policy. The policymaker may at this point want to add their own way of weighing the different agents' outcomes together—by applying a specific social welfare function—or not. A commonly adopted social welfare function in the literature is the equally-weighted utilitarian function: one where the present-value utility functions of all agents at time zero is just summed up. Such a social welfare function is normative—the particular function is one that implicitly puts a high weight on equality—and not cannot generally be justified otherwise.[37,38]

The nature of contracts The heterogeneous-agent literature, where wealth inequality is a natural outcome, fundamentally builds on incomplete markets. The canonical model has no insurance but riskless saving, along with an exogenous borrowing limit, with the loose empirical motivation that many risks in life appear uninsurable, except via saving and only limited borrowing. Yet a number of insurance markets do exist, life insurance, property insurance, and health insurance being prominent examples. In addition, most countries have some degree of social security and publicly provided health services at low cost; also note that in some countries, such as the U.S., there is personal bankruptcy protection, providing further insurance. Insurance against professional failure (captured by "wage risk" in the model), or divorce, are examples where the complete lack of formal insurance markets seems a more appropriate assumption.[39]

Given all this, what is appropriate modeling of the nature of contracts? The literature tends to adopt assumptions that are easy to implement and seem like reasonable descriptions of reality. Many models include a description of the publicly provided safety net. Examples of important and non-trivial extensions to the canonical model can be found in the literature that incorporates personal bankruptcy protection into the canonical setting, as pioneered by Chatterjee, Corbae, and Rios-Rull (2008) and Livshits, MacGee, and Tertilt (2007), and models how publicly available information on individuals (e.g., credit scores) is used as a determinant of borrowing contracts, such as in Chatterjee, Corbae, Dempsey, and Ríos-Rull

[37]Equal weights will want the planner to distribute consumption so as to equalize the marginal utilities of all agents. When utility is additive in consumption, this implies equal consumption for all agents.

[38]The equal-weight utilitarian welfare function, applied at time 0 for present-value utility, can be justified normatively if, at time zero, all agents are in identical situations—but may differ ex post due to shocks. This measure is referred to as one "behind the veil of ignorance", a notion introduced in Rawls (1971).

[39]Informal insurance markets, through family and social networks, may still be active.

(2023). Common to all these frameworks is a difficulty of conducting policy analysis: one would expect the nature of contracts to react to any policy change, and if the model does not endogenize the source of market incompleteness assumed, it is vulnerable to a Lucas critique.

To be concrete, in a canonical model it seems model-feasible to enact a policy that entirely insures against wage/earnings risk, and such a policy would improve welfare in an ex-ante sense.[40] If enacted in the real world, it would likely lower average income significantly, as efforts and investments in human capital would likely be significantly reduced: after all, such a policy would be "socialism pure", a system which few economists would argue in favor of. Most likely, the reason why markets do not provide insurance for earnings shocks is a significant degree of private information leading to moral hazard and adverse selection, as well as limited ability of consumers to commit.[41] One approach would thus be to formulate models with explicit information/commitment frictions and derive the optimal contract given these frictions; such models would be "Lucas-proof". This approach is technically challenging, however, and often lead to contract types that do not resemble what we observe in reality.[42] Thus, there remains a tension between descriptive accuracy and Lucas-proof modeling.

21.3 Reasons why inequality matters for aggregates

The previous sections described inequality, along with a number of theories of different dimensions of inequality. Here we briefly address how the presence of inequality affects macroeconomic aggregates in important ways.

21.3.1 Long-run channels

We only briefly mention the channels through which inequality is important for growth and development. The brief mention is not for lack of importance, however; since long, many connections between inequality and growth/development have been studied in the macroeconomic literature. The discussion here will mainly be qualitative.

Incentives

There is a view that inequality is "good": not by itself, but because it may reflect the presence of incentives to work hard and to innovate and accumulate. Tax and transfer systems aimed at equalizing consumption would then normally generate lower aggregate output, and likely worsen welfare for a large set of agents. Take, for example, an Aiyagari model with variable labor supply. A system that taxes earnings and capital income at proportional rates while transferring the revenue in lump-sum, equal amounts to all agents would indeed make the distributions of disposable income and consumption less dispersed.

[40] That is, suppose agents are all identical at time zero, with equal wealth and without having experienced any shocks yet. Then they are all identical and would benefit from full insurance.

[41] Mirrlees (1971) and a large follow-up literature studies optimal insurance contracts in private-information economies. Note also that publicly supported bankruptcy regulation is often an imperfect solution to a commitment problem; see Mateos-Planas, McCrary, Ríos-Rull, and Wicht (2025).

[42] See Allen (1985) and Cole and Kocherlakota (2001) for progress in this regard.

(Formulate such a model and use our program package to solve it for a steady state and you will see!) However, such a system would generate a lower capital stock in steady state as well as fewer hours worked. On average, people will most likely be worse off under such a system, even as measured from time 0, taking the transition into account. The qualification "most likely" is necessary since redistribution, even when distortionary, can improve welfare when markets for insurance are missing: consumption smoothing across states is desirable but markets do not, by assumption in this case, allow it.[43] Of course, ideally, then, a policy that improves on market performance, if such a policy is feasible, would be desirable. However, deeper frictions, such as private information (moral hazard or adverse selection) or lack of commitment, might make such improvements impossible.[44] I.e., no government policy would then allow us to achieve better outcomes.

A recent literature looks at the incentives for innovation and patent creation in rich countries. Can significant inequality hamper these activities? To the extent insurance markets are poor and the potential innovators, whose projects involve risks, cannot insure against these risks, they may benefit from social insurance schemes or from joining larger companies where the payoffs are somewhat smoothed out across states. How and where—in what market forms—innovation activities occur is therefore an important element behind productivity growth. Arguably, in many rich countries the need for insurance may be limited. The outgrowth of unicorn companies and the remarkable wealth accumulation associated to them is a phenomenon that is not just observed in countries like the United States: they are observed, along with high Gini coefficients for wealth, in Scandinavian countries as well, where the social insurance schemes are much more developed. It is conceivable, however, that the even more extreme wealth accumulation outcomes observed in the United States can be a result of lower government involvement and freer markets, reflecting incentive effects that generate more innovative activity.

Credit-market restrictions and inequality traps

A basic, and very important, issue concerns human capital accumulation and how inequality may hinder it.[45] Education is to some extent a consumption good but the consensus in the literature is still that human capital is key for production and economic development. The phenomenal developments within AI recently is an example of technology developments that surely would be impossible without highly educated innovators. The "access to education", which clearly is a prerequisite for advanced development of an economy, can be interpreted as "education involves a fixed cost". I.e., if there are no schools in one's country, or one's neighborhood, or if there are but they involve significant tuition fees, then one's wealth level becomes a key determinant of the possibility of attending school/university, unless it is possible to borrow to finance the education. Suppose a worker can work as unskilled and earn a present-value lifetime labor income of w_u or attain education and obtain the skilled lifetime income $w_s > w_u$, and suppose the education costs F. Then if $w_s - w_u > F$, it would pay off for the individual to obtain education. They would thus choose to become educated,

[43] A policy that is not distortionary but that distributes from the lucky, i.e., those with high earnings or wage shocks, to the unlucky, would of course be even better.

[44] That is, even the distortionary tax system just described may not be feasible.

[45] For an early treatment, see Galor and Tsiddon (1997).

unless the cost of education is paid upfront and the income gains materialize later and the individual cannot self-finance education or borrow to attain it. Without any possibility of borrowing, their own assets would need to be at least above F for education to be a feasible choice. Thus, the fraction of individuals who would choose education given these income levels equals the fraction of people with wealth above F in the wealth distribution; hence, the wealth distribution matters for educational attainment. A full general-equilibrium treatment of this idea would endogenize the earnings w_u and w_s, perhaps with an aggregate production function like that studied earlier in this chapter; in such a setting, the fraction of people who obtain education along with the earnings distribution would be affected by the initial wealth distribution. Poor individuals can also end up in "trapped dynasties" where the next generation will not obtain an education either, due to the low earnings of the present generation. On the level of an entire economy, there can potentially also be a trap, with low overall GDP due to credit constraints preventing a large part of the population from becoming educated. To what extent this mechanism is an important factor behind inequality within a country, or behind the differences in productivity across countries, are still open issues.

21.3.2 The business cycle

The role of inequality for understanding business cycles has received significant attention in macroeconomic research over the last decades. In particular, heterogeneous-agent models—essentially like those discussed at several points in the book and in detail in the previous section of this chapter—with one or more sources of aggregate fluctuations, including with nominal frictions, are now viewed to be an important part of our macroeconomic toolkit. A key reason why this literature has had impact in macroeconomics is that it allows the marginal propensity of consumption, MPC, to rise relative to that coming out of a representative-agent, dynastic model. In the latter, recall that the permanent income effects of transfers are very small (on the order of the interest rates); in heterogeneous-agent models many consumers have very high MPCs (those who are literally borrowing-constrained have a marginal propensity of 1).[46] With larger MPCs, transfer policies, such as typical fiscal policy, will have larger effects on overall demand and hence make such stabilization policy more powerful.

The how-to

In principle, an endogenous wealth distribution is a high-dimensional study object, though as we have seen earlier in this chapter, it is straightforward to solve for the distribution in steady state. In particular, in the basic Aiyagari model, it suffices to guess on a steady-state capital stock and then, given this guess, be able to solve a dynamic programming problem with one endogenous state (asset holdings, or cash on hand), find the stationary distribution of asset holdings generated by the implied decision rules and iterate until the sum of asset holdings in the stationary distribution matches the capital stock assumed. This numerical procedure finds a solution within seconds. However, to study aggregate uncertainty requires also determining how the distribution evolves stochastically over time. In principle, the

[46]The Keynesian consumption function, in particular, was often estimated to feature MPCs of 0.5 or higher.

distribution of wealth will affect prices, so each agent's dynamic programming must have as a state variable the distribution of wealth, which is a high-dimensional object: solving dynamic programming problems is explosively more costly as more states are included. The agent also needs to know the mapping from the wealth distribution to prices. Overall, it simply seems infeasible to study such economies, even with fast computers.

The literature has addressed this in two ways. One has been to lower the ambition level in terms of heterogeneity and assume that there are two, or a very small number, of representative agents in the economy. The leading, simplest example of this approach would have one agent who is a worker and cannot save. Hence, this agent is "hand-to-mouth", i.e., consumes all current income and thus has an MPC of 1. The other agent, then, would be a pure "rentier" who does not work but simply owns and rents out capital and consumes from the capital income. Such an agent would have an MPC much closer to zero, as in a standard representative-agent case. The fraction of consumers of each type can then be selected so as to obtain an aggregate MPC of, say, 0.5. Such a version of a New Keynesian setting, often referred to as a "TANK" (Two-Agent NK) model, is used at many policy institution.

The other approach in the literature has been to confront the computational challenge head on. Krusell and Smith (1998) showed one path forward, and the following decades have added numerous elaborations and alternatives, with the result that methods for solving heterogeneous-agent models with aggregate shocks are now taught in graduate programs and considered standard second-year material, and they also used at policy institutions. One method that is particularly accessible conceptually and computationally relies on a presumption that, around a steady state, the behavior of aggregates can be well approximated as a linear function of the shocks hitting the economy.[47] The procedure boils down to computing the transitional deterministic behavior of the economy in response to a one-time unexpected shock of a given size; then random, repeated shocks of different sizes can be computed using linearity.[48] Further improvements are being added at a rapid rate, including for cases where second-moment effects and nonlinearity play a central role.

Insights and relevance

As already stated, the heterogeneous-agent approach to business cycles allows us to derive MPCs that are more in line with data. Consider for example an Aiyagari-style model, augmented so that it matches the wealth distribution in steady state. Then there will be MPCs within the distribution of agents ranging from 0.01 or so to 1, and depending on the exact form of the distribution, the average MPC will be 0.1 or quite a bit higher. These features are captured in the stylized Figure 21.9 depicting the decision rule for saving.

A particular recent version of Aiyagari models also includes the so-called "wealthy hand-to-mouth" agents. These are individuals who, despite being relatively wealthy, have high MPCs due to their asset portfolios being rather illiquid.[49] A particularly simple, ad hoc way

[47] See Boppart et al. (2018) and Auclert, Bardóczy, Rognlie, and Straub (2021).

[48] Linearity means that responses can be scaled by shock size and added up, including for a vector of different kinds of shocks. It also means that the responses to shocks will be identical whether they are random or deterministic: recall from Chapter 7 that certainty equivalence applies. Hence also the effects of random shocks can just be added up.

[49] See Kaplan and Violante (2010).

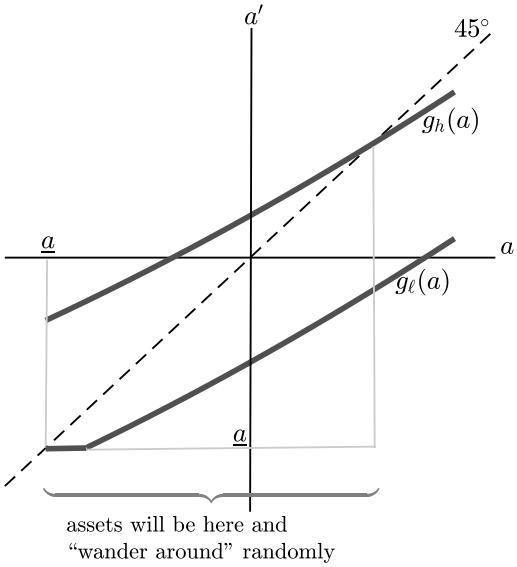


Figure 21.9: MPCs high for low asset values (1 when constraint binds).

of generating this mechanism is to assume that, independently of the agent's current asset position a, there is an iid random event whereby $a' \geq a$ is imposed: i.e., if the event occurs, the agent cannot decumulate assets. Thus, in case this event occurs in a period where the non-asset income is low and the agent would like to draw down on asset holdings in order to consume, the agent will be constrained and have an MPC equal to 1. With this additional mechanism, obviously the average MPC rises even further.

As a result of higher average MPCs, fluctuations in aggregate consumption are brought closer to the data: the correlation between consumption and output rises and the excess sensitivity puzzle (see Chapter 11) can be explained: aggregate income shocks, even recent ones, move consumption directly through their effects on borrowing-constrained agents. However, the barebones Aiyagari model, even with an added mechanism that generates MPC heterogeneity, cannot contribute much to increased output movements, because the "demand side" of the model is rudimentary: output is given by the production function, where a productivity shock is the main reason for fluctuations. Models where demand plays a central role in determining output, at least in the short run, are thus needed. One example is to be found in the literature combining the Aiyagari-style model with New Keynesian frictions (so-called HANK models). Let us, however, look at an another very simple example where demand matters, similar to that alluded to in Chapter 3.5.1. Suppose output is given by $\hat{A}_t k_t^\alpha$, that is, labor is not a variable factor, but where $\hat{A}_t = A_t c_t^\omega$, where A_t is exogenous TFP and c_t is consumption at time t, with $\omega \geq 0$: there is a positive externality to consumption. That is, if people demand higher consumption, output rises, everything else equal: output is, to this extent, demand-determined.[50] Suppose the model is otherwise neoclassical, of the Aiyagari (1994) variety, but with the addition of a wealthy-hand-to-mouth mechanism as the one just discussed above: for each agent, the probability that saving is not allowed

[50]For models of this sort, see Krueger et al. (2016) and for more elaborate frameworks generating reduced forms similar to this framework, see Bai, Ríos-Rull, and Storesletten (2025) and Huo and Ríos-Rull (2015).

to decrease between t and $t+1$, γ, is iid. Figure 21.10 shows the results of the impact response of consumption and output of a one-percent shock to the exogenous part of TFP. If $\omega = \gamma = 0$, output (in the right-hand-side panel) rises mechanically by 1 percent, whereas consumption rises by 0.3 percent. If γ is high (0.6), so that many agents have a high MPC, consumption rises by over 1.1 percent, but output is unaffected: investment responds negatively, as total resources available for consumption and investment is fixed. However, with a significant externality, the strong consumption response also generates a significant rise in output. We see that the two mechanisms—high MPCs and a demand channel—reinforce each other and generate strong propagation from TFP shocks, even in the complete absence of an endogenous hours channel.

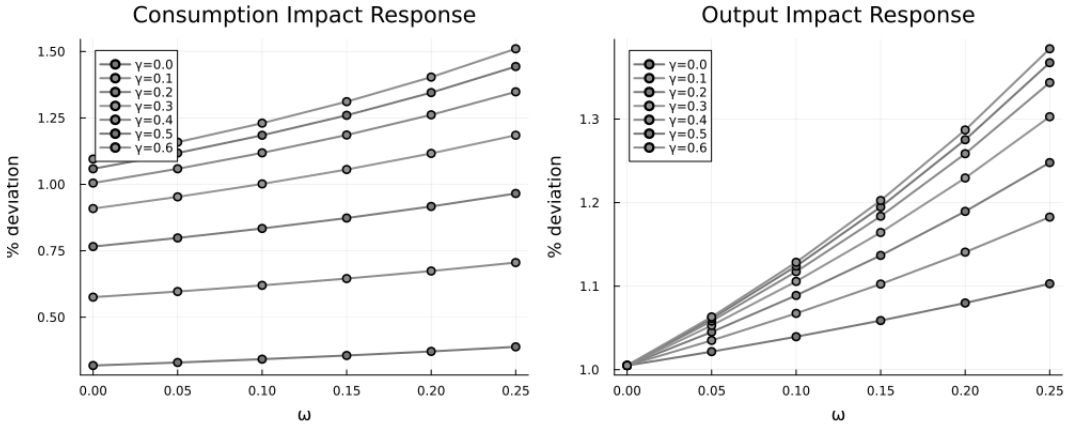


Figure 21.10: Impact responses to a 1 percent TFP responses).

A further insight is that the effects of government policy, such as fiscal transfers or monetary policy, depend on which population groups are affected. In order to boost demand, for example, effective policy instruments are those that target high MPC consumers. Moreover, the marginal propensities to work and invest in risky vs. non-risky assets will also differ across the population; i.e., policymakers more generally must assess the distribution of marginal propensities when comparing different policy options.

Models with aggregate fluctuations and heterogeneous agents also allow policymakers to look at the differential impacts of policy on the welfare of different agents. Traditional macroeconomic models have only allowed us to look at the unemployment dimensions of policy—how effects on the employed differ from those of the unemployed—but unemployment is of course only one determinant of inequality, with effects through the full wage distribution playing an arguably important role. Thus, models with rich, endogenous wage inequality accompanied by wealth inequality and less than full insurance also offer many new avenues for policy analysis.

21.3.3 From micro to macro: more heterogeneity

The previous two subsections briefly discuss how macroeconomic aggregates may behave differently, in the long as well as in the short run, due to the presence of inequality in incomes and wealth. Let us finally mention a number of important lines of macroeconomic research that further explore heterogeneity, of different kinds, on the microeconomic level.

All macroeconomic models fall short of modeling the full complexity at the microeconomic level and there is always a question of whether this abstraction is distorting our analysis. One motivation behind this kind of research is a desire to check that conclusions from our are robust when we allow for the main sources of difference across households. As an example, we have just seen that incorporating inequality in incomes and wealth on the microeconomic level will change our standard dynastic model in ways that can be important in addressing some questions, such as stabilization policy.

Age differences between individuals have already been discussed in our text, as a natural element of overlapping-generations models. The age structure of an economy can matter for aggregate saving, but also when studying social security and pensions. Currently, fertility is low in many countries, which means that an increasing burden will be placed in the future on the working individuals to provide for the elderly: aggregate labor supply will be significantly affected. Saving and labor supply are also affected by the household structure. Taking marriage and cohabitation into account can be important, for example because the response of labor supply to transfers and to changes in labor market conditions can become strong for "second earners" in a household.[51] In addition, cohabitation appears to be countercyclical—presumably to cut costs when incomes are low—and have a downward trend.

A large literature explores labor markets from a macroeconomic perspective and there, heterogeneity in many dimensions, such as individual's education, occupation, work experience, health status, or location of residence, is often incorporated into the analysis. Sometimes, such models involve search and matching markets (as in Chapter 20), they pose additional computational challenges: the current distribution of agents across matches becomes relevant for any given individual and this distribution can sometimes be captured by a simple statistic, such as labor market tightness, but sometimes it cannot.

Heterogeneous-agent macroeconomics is expanding rapidly, in part because computational power and methods are advancing at a rapid pace, allowing our macroeconomic analyses to benefit from detailed microeconomic foundations. Finally, more realistic microeconomic foundations allow us to connect the model with microeconomic data and microeconomic studies. By relating the model to a wider range of data, we can bring more information to bear on our research questions.

[51] Married women have been documented to have a more elastic labor supply and a less strong attachment to the labor force.

Chapter 22

Heterogeneous firms

Toshihiko Mukoyama

22.1 Introduction

In most macroeconomic models (and earlier in this textbook), it is assumed that there exists an aggregate production function

$$Y = F(K, L),$$

where Y is the aggregate output, K is the aggregate capital, and L is the aggregate labor. This assumption is justified if the production functions for all firms are homogeneous. In reality, firms are heterogeneous in many dimensions. There are large and small firms, young and old firms, growing and contracting firms, and productive and unproductive firms.

Answering many macroeconomic questions requires explicitly taking firm heterogeneity into account. For example, how should we encourage (or discourage) the entry of new firms? How should we support growing firms? Should we be concerned about the growing prominence of "mega-firms"? What are the causes and consequences of the decline in firm entry and reallocation of resources across firms? Analysis based on the aggregate production function cannot answer these questions. It may also be the case that some economic policies may have different effects on the macroeconomy if firms have different degrees of heterogeneity.

In this chapter, we consider such questions. Looking at the data, we will see some indications that the prominence of large firms in the U.S. economy has been rising in recent years. The reallocation of resources through the entry and exit of firms, as well as the expansion or contraction of firms, seems to be slowing down in recent years (often referred to as the "decline in business dynamism"). These phenomena have potentially important consequences in the macroeconomic context. The rise of big firms may lead to an increase in their market power. The market power in the product and the labor market could be linked to market distortions and changes in the labor share. The lack of reallocation of resources from unproductive firms to productive firms may lead to lower aggregate productivity due to "misallocation" (resource allocation that is suboptimal). Misallocation may also lead to a slower rate of innovation and aggregate productivity growth. The dominance of large firms may also have implications for other macroeconomic phenomena, such as business cycle fluctuations.

To answer these questions, we need to break out of the aggregate production function. This chapter covers the basic facts, models, and methods for analyzing an economy with

heterogeneous firms.

22.2 A simple model

We begin by considering a simple example where firm heterogeneity matters for macroeconomic analysis.[1] Suppose that there is a unit mass of firms. The firms produce a homogeneous good under perfect competition. The production function of firm i (where i is the index of firms: $i \in [0,1]$) is

$$y_i = a_i F(\mathbf{x}_i)^\gamma,$$

where y_i is the output of firm i and $\gamma \in (0,1)$. Firms are heterogeneous in their productivity; a_i is different across firms. $\mathbf{x}_i$ is the input vector for firm i. Assume that $F(\mathbf{x}_i)$ exhibits constant returns to scale. Then, because $\gamma < 1$, the overall production of y_i exhibits decreasing returns to scale in inputs $\mathbf{x}_i$. The decreasing returns property is important. With constant returns, the firm(s) with the largest a_i takes over the entire production of the economy, and the outcome is either (i) a monopoly or oligopoly of one or a few firms, which would contradict the perfect-competition assumption; or (ii) only the most efficient firms with common a_i operate as price takers, which would return to the homogeneous-firms scenario. Let $\mathbf{X}$ be the endowment vector of inputs in the economy.

Due to the constant-returns property of $F(\mathbf{x}_i)$, we can solve the firm's problem in two steps: first, solve the cost-minimizing combination of inputs for one unit of $F(\mathbf{x})$. Second, decide the optimal scale of production. The first stage is common across firms:

$$\min_{\mathbf{x}} \ \mathbf{p}\mathbf{x}$$

subject to

$$F(\mathbf{x}) = 1,$$

where $\mathbf{p}$ is the vector of input prices. Let the solution of this problem be $\mathbf{x}^*$ and the minimized unit cost be $c \equiv \mathbf{p}\mathbf{x}^*$.

Let $m_i = F(\mathbf{x}_i)$ be the choice of the firm i's combined inputs. The constant-returns property implies that the optimal input choice is $\mathbf{x}_i = m_i \mathbf{x}^*$ and the cost of production is cm_i. The second stage optimization problem is

$$\max_{m_i} \ a_i m_i^\gamma - c m_i. \tag{22.1}$$

The first-order condition for this problem is

$$a_i m_i^{\gamma-1} = \frac{c}{\gamma}. \tag{22.2}$$

Therefore, $y_i = (c/\gamma) m_i$ for all i. Adding up for all i,

$$Y = \frac{c}{\gamma} M \tag{22.3}$$

[1] A similar framework is used by Hopenhayn (2014a). Some of the results below overlap with his.

holds, where

$$Y = \int y_i di \tag{22.4}$$

is the total output and

$$M = \int m_i di. \tag{22.5}$$

Note that, in equilibrium, $M = \int F(\mathbf{x}_i)di = F(\mathbf{X})$ has to hold. Let us define

$$A \equiv \left(\int a_i^{\frac{1}{1-\gamma}} di \right)^{1-\gamma}. \tag{22.6}$$

From (22.2),

$$A = \frac{c}{\gamma} M^{1-\gamma}$$

holds. Combining with (22.3) and $M = F(\mathbf{X})$,

$$Y = AF(\mathbf{X})^{\gamma}. \tag{22.7}$$

In this environment, this relationship can be viewed as the aggregate production function.[2] The heterogeneity of firms matter through the aggregation (22.6): the aggregate outcome is influenced by the distribution of a_i to the extent that it yields different values of A in (22.6).

To illustrate, suppose that a_i follows a lognormal distribution, where $\log(a_i) \sim N(\nu - \sigma^2/2, \sigma^2)$. From the property of the lognormal distribution, the average of a_i, $\int a_i di$, is $\exp(\nu)$. However, it can be computed that[3]

$$A = \exp\left(\nu + \frac{\gamma}{1-\gamma} \frac{1}{2} \sigma^2 \right). \tag{22.8}$$

Therefore, the dispersion parameter σ influences the level of A even when the average productivity $\exp(\nu)$ is constant. This result holds because the productive resources are endogenously allocated: a productive firm uses more input than an unproductive firm and therefore has a greater presence in aggregate production than merely having higher productivity. When the dispersion parameter σ is larger, the economy has more room to allocate resources to the highly productive firms in the right tail. Allocation of inputs is the key to analyzing heterogeneous firms: when the inputs are not allocated optimally, aggregate productivity can be less than what can be achieved optimally. In this chapter, we always keep two questions in mind: (i) how the distribution of a_i is determined, and (ii) how the economy allocates resources to different firms.

[2] An example of this aggregation is when the production function is $y_i = a_i(k_i^{\alpha} \ell_i^{1-\alpha})^{\gamma}$, where k_i is firm i's capital input, ℓ_i is the firm i's labor input, and $\alpha \in (0,1)$. In this case, the aggregate production function is

$$Y = A(K^{\alpha} L^{1-\alpha})^{\gamma},$$

where A is given by (22.6). The rental rate of capital in equilibrium is $r = \gamma\alpha A(K^{\alpha}L^{1-\alpha})^{\gamma-1}K^{\alpha-1}L^{1-\alpha}$, the wage rate is $w = \gamma(1-\alpha)A(K^{\alpha}L^{1-\alpha})^{\gamma-1}K^{\alpha}L^{-\alpha}$, and the unit cost of production is $c = (r/\alpha)^{\alpha}(w/(1-\alpha))^{1-\alpha}$.

[3] See Appendix 22.A.1 for derivation.

22.3 Firm heterogeneity in the data

This section describes some facts related to firm heterogeneity. We will focus on the U.S. data. The statistics presented here are based on publicly available data.[4] The first natural question is: how heterogeneous are the U.S. firms? Figure 22.1 shows the firm size distribution as the number of firms in each size category, as a fraction of the total number of firms. Here, the firm size is measured by the number of employees.

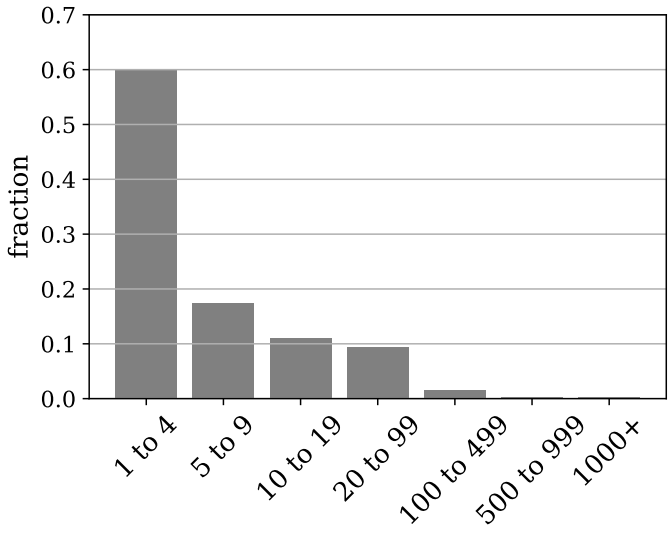


Figure 22.1: Distribution of firm size in 2022.

Source: Business Dynamics Statistics.

Figure 22.1 shows that the firm size distribution is quite dispersed. There are over 5 million firms in the U.S., and the majority are very small firms with 1 to 4 employees. At the same time, there are over 10,000 large firms with more than 1,000 employees, as well as over 1,000 firms with 10,000 employees or more.

The fact that very small firms account for the majority of firms does not imply that large firms are unimportant. Figure 22.2 plots the employment share of each size category. Approximately half of all employees work at firms with 1,000 or more employees. In fact, approximately 30% of workers are employed by very large firms with 10,000 or more employees.

Firm dynamics literature often uses data at the establishment level. An establishment is a fixed physical location where economic activity occurs; it is more straightforward to identify an establishment than a firm. A firm is a collection of establishments under common ownership, and it is often difficult to identify a firm in an administrative dataset. Establishments are also heterogeneous. Figure 22.3 is the establishment size distribution. There are over 7 million establishments in the U.S. economy, and approximately half are very small establishments with 1 to 4 employees.

Figure 22.4 plots the number of establishments that are owned by each firm size cate-

[4] All figures in this section are drawn from the U.S. Census Bureau's Business Dynamics Statistics. See `https://bds.explorer.ces.census.gov/`.

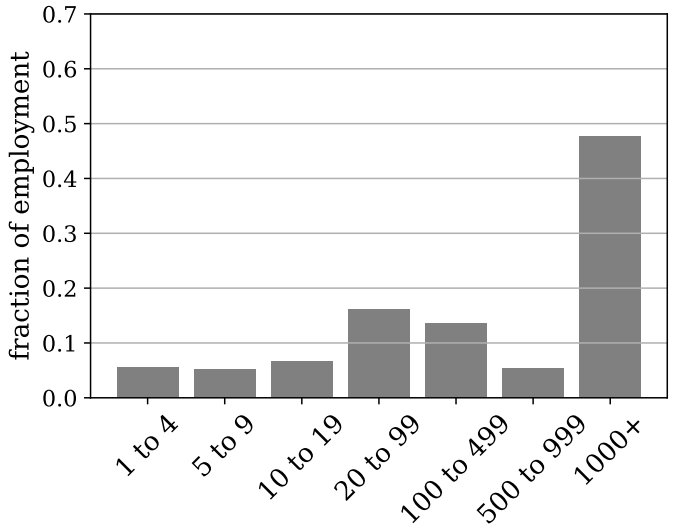


Figure 22.2: Employment share of each size category in 2022.

Source: Business Dynamics Statistics.

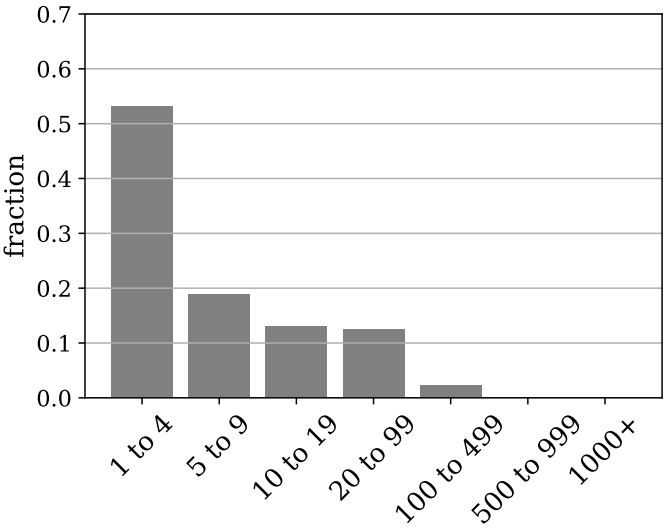


Figure 22.3: Distribution of establishment size in 2022.

Source: Business Dynamics Statistics.

gory. It shows that many establishments are owned by large firms (approximately 16% of all establishments are owned by firms in the 1,000+ category). Whereas almost all "1 to 4" category firms own only one establishment, the firms in the 1,000+ category own 100 establishments on average. Very large firms with 10,000 or more employees own approximately 600 establishments on average.[5]

Aside from the cross-sectional heterogeneity, U.S. firms conduct significant adjustments

[5]See Cao, Hyatt, Mukoyama, and Sager (2022) for a detailed analysis of the number of establishments per firm and its time-series properties.

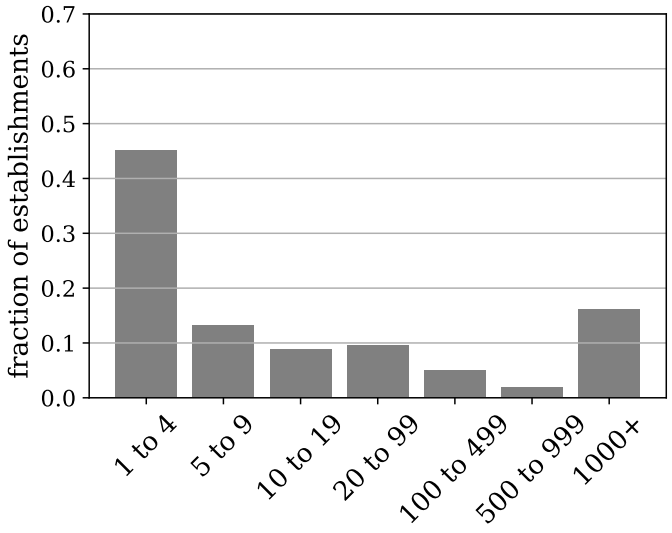


Figure 22.4: Fraction of establishments owned by each firm size category in 2022.

Source: Business Dynamics Statistics.

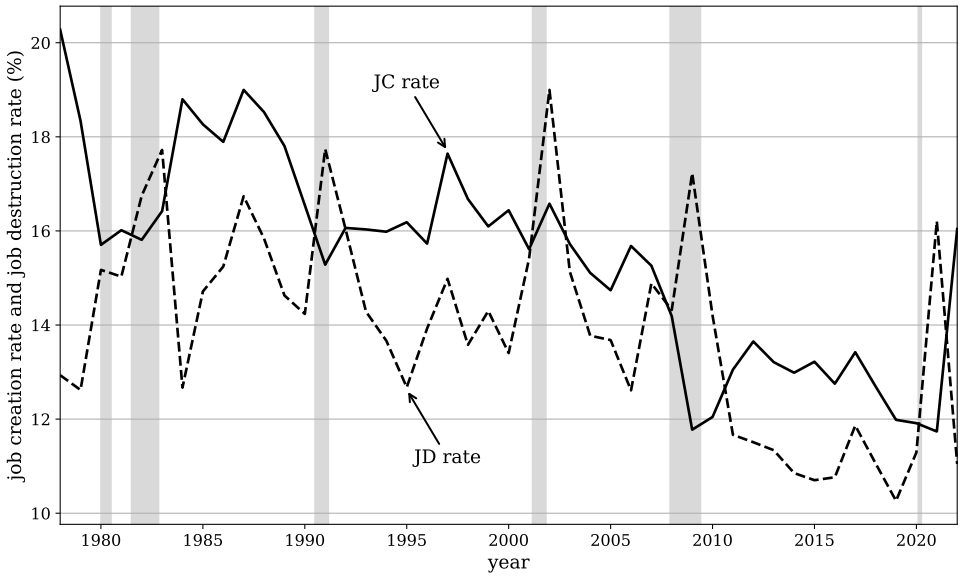


Figure 22.5: Annual job creation and destruction rates (establishment level).

Source: Business Dynamics Statistics.

over time. One measure of a firm's size adjustment is job creation and destruction. Job creation (JC) refers to the expansion of firms or establishments, whereas job destruction (JD) refers to the contraction of firms or establishments. BDS publishes the establishment-

level JC and JD rates. The JC rate is defined as:

$$JC_t \equiv \frac{\sum_{i:\ell_{it}>\ell_{i,t-1}}(\ell_{it}-\ell_{i,t-1})}{\bar{L}_t}, \tag{22.9}$$

where ℓ_{it} is the employment of establishment i at year t, L_t is the total employment at year t (which is the sum of ℓ_{it}), $\bar{L}_t \equiv (L_t + L_{t-1})/2$. In words, the JC rate is the sum of employment increases in all expanding establishments, divided by the total employment (the average of time t and $t-1$). The JD rate is similarly defined as:

$$JD_t \equiv \frac{\sum_{i:\ell_{it}<\ell_{i,t-1}}(\ell_{i,t-1}-\ell_{it})}{\bar{L}_t}.$$

The JD rate is the sum of the employment decrease by contracting establishments, divided by the total employment (the average of time t and $t-1$). JC and JD, often called gross job flows, measure the magnitude of labor reallocation across establishments. Figure 22.5 plots the JC and JD rates from the BDS dataset. The shaded area is the recession period defined by the National Bureau of Economic Research (NBER).[6] Three properties are notable. First, the magnitude of JC and JD is large. Both the JC and JD rates exceed 10% in any given year. Second, both rates are cyclical. When a recession arrives, the JC rate declines and the JD rate increases. Third, there is a general declining trend in both the JC and JD rates. It is known that a wide range of indicators of reallocation, including the JC and JD rates, have declined in recent years. Some researchers call this trend the "declining business dynamism" of the U.S. economy.

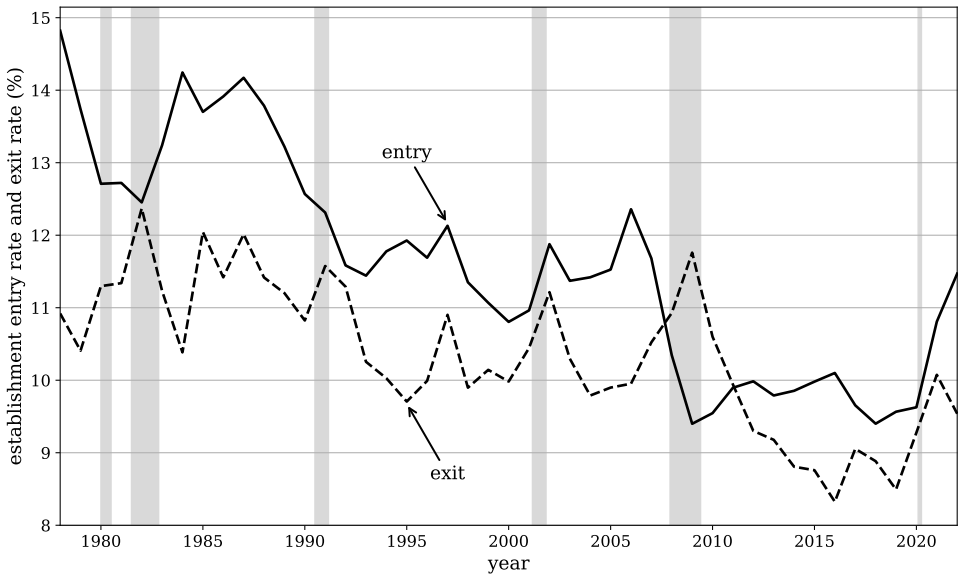


Figure 22.6: Annual establishment entry and exit rates.

Source: Business Dynamics Statistics.

Another measure of reallocation is the rate of entry and exit. Many firms and establishments enter and exit every year. Figure 22.6 plots the entry rate and exit rates of

[6] See `https://www.nber.org/research/business-cycle-dating`.

establishments. The entry rate is defined as the number of entering establishments between $t-1$ and t divided by the total number of establishments (the average of time $t-1$ and t). The exit rate is defined as the number of exiting establishments between $t-1$ and t divided by the total number of establishments (the average of time $t-1$ and t). One can observe similar properties here as in the JC and JD rates: the entry and exit rates are large, cyclical, and there are overall declining trends.

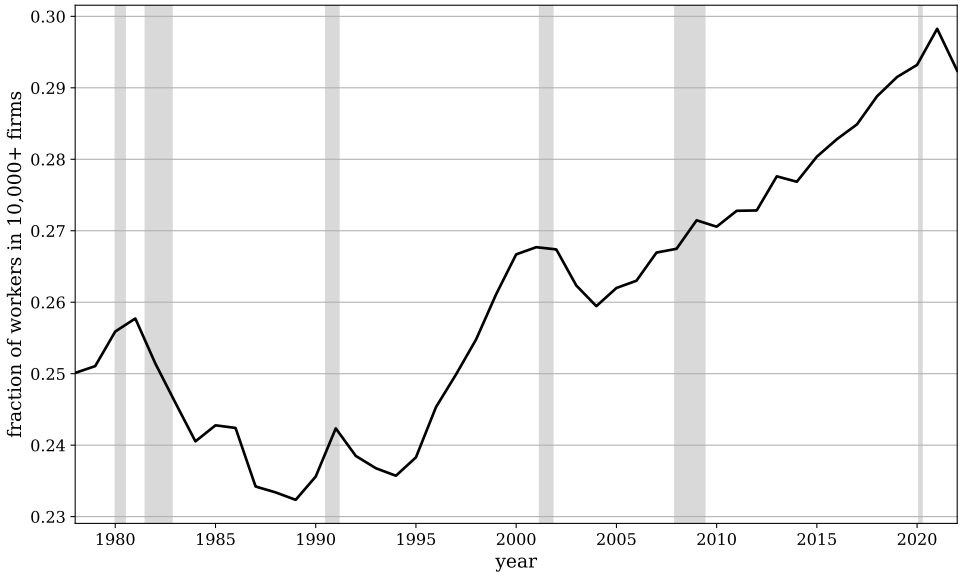


Figure 22.7: Fraction of employees working at 10,000+ employee firms.

Source: Business Dynamics Statistics.

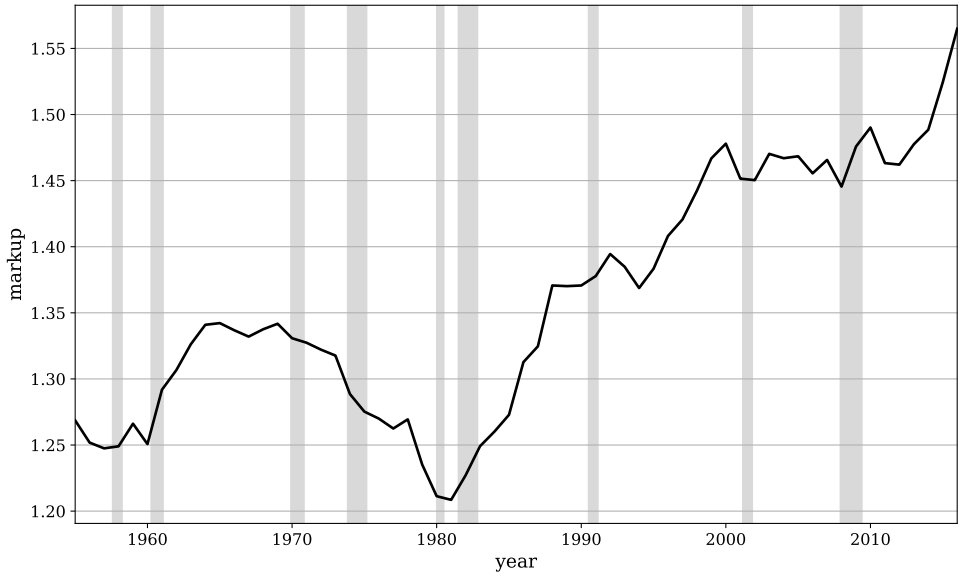


Figure 22.8: Markup of the U.S. public firms.

Source: De Loecker et al. (2020).

Over the last few decades, there have been significant changes in the heterogeneity among U.S. firms. In addition to the "declining dynamism" described above, one topic that caught researchers' attention is the dominance of large firms. Figure 22.7 plots the fraction of workers employed by firms in the 10,000+ size category.[7] This fraction has steadily increased since the early 1990s, indicating that large firms are starting to dominate the U.S. economy. This dominance raised concerns about the market power of large firms, and is consistent with another strand of research that tries to measure the trend of market power in the U.S. economy. Figure 22.8 reproduces Figure 1 of De Loecker et al. (2020). It measures the trend of the average markup (i.e., price over the marginal cost) of the U.S. public firms in the Compustat dataset. The markup series exhibits an increasing trend since the 1980s.[8]

22.4 Reallocation and misallocation

The previous section shows that there is a large degree of reallocation among U.S. firms. How much does the reallocation matter for aggregate productivity? In Section 22.1, we have seen that firm heterogeneity affects the aggregate outcome through the endogenous allocation of inputs. In a dynamic economy where firm productivity changes over time, one can imagine that the constant reallocation of inputs can have an important impact on aggregate productivity.

Foster, Haltiwanger, and Krizan (2001) illustrate the quantitative impact of reallocation through the following simple accounting framework. Let us denote the productivity of establishment i (they use establishment-level data and not firm-level data) at time t as a_{it}. The (output-weighted) average productivity $\bar{A}_t$ is defined as

$$\bar{A}_t \equiv s_{it} a_{it},$$

where s_{it} is the output share of establishment i. Then, by denoting the $x_t - x_{t-1}$by Δx_t,

$$\begin{aligned}\Delta \bar{A}_t = & \sum_{i \in C} s_{it-1} \Delta a_{it} + \sum_{i \in C} (a_{it-1} - \bar{A}_{t-1}) \Delta s_{it} + \sum_{i \in C} \Delta a_{it} \Delta s_{it} \\ & + \sum_{i \in N} s_{it} (a_{it} - \bar{A}_{t-1}) - \sum_{i \in X} s_{it-1} (a_{it-1} - \bar{A}_{t-1})\end{aligned}$$

holds, where C is the set of continuing establishments (establishments that exist in both time $t-1$ and t), N is the set of new establishments (establishments that enter between time $t-1$ and t), and X is the set of exiting establishments (establishments that exit between time $t-1$ and t). The increase in average productivity can occur for five distinct reasons. First, each of the existing establishments can increase its productivity. Second, an establishment with higher-than-average productivity can increase its market share. Third, the first two effects can be magnified if both occur at the same time (i.e., a high-productivity establishment

[7] In drawing this figure, the distinction between firms and establishments is very important. See Appendix 22.A.2.

[8] The evolution of market power, both in the product market and the factor market, remains an active research topic. The studies that follow De Loecker et al. (2020) highlight important methodological limitations and industry heterogeneity. Useful surveys include Miller (2025) and Syverson (2025).

raises the share and its own productivity). Fourth, the entering establishment can be better than the average. Fifth, the exiting establishment can be worse than the average. All factors except for the first one can be interpreted as the contribution of reallocation. That is, if $\Delta s_{it} = 0$ and there are no entry and exit, the only way for the aggregate productivity to increase is for each establishment to increase its productivity. Using the U.S. Manufacturing data from 1977 to 1987, Foster et al. (2001) estimate (see their Table 8.4) that the aggregate change in multifactor productivity (the change in output that is not accounted for by the change in capital, labor, and intermediate goods) is 45% accounted for by the first factor. The remaining 55% is the contribution of reallocation. This decomposition highlights the importance of reallocation in determining aggregate productivity growth.

Recently, a large body of literature has evaluated the role of various frictions that hinder the optimal allocation of resources. This literature emphasizes the existence of the *misallocation* of productive inputs as the source of low aggregate total factor productivity. A subset of literature, such as Restuccia and Rogerson (2008) and Hsieh and Klenow (2009), emphasizes firm-specific distortions as the sources of misallocation. To see how firm-specific distortions can affect aggregate productivity, consider the model of Section 22.1 and add an assumption that the government taxes the output of firm i at the rate of τ_i. Thus, instead of the problem (22.1), the firm solves

$$\max_{m_i} (1-\tau_i) a_i m_i^{\gamma} - c m_i.$$

The rest of the model is the same, and the GDP is still measured as $Y = \int y_i di$. After going through similar steps as in Section 22.1, one can show that the aggregate production function still takes the form of (22.7), but A is modified to

$$A = \frac{\int a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}} di}{\left(\int a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{1}{1-\gamma}} di \right)^{\gamma}}. \tag{22.10}$$

One can easily see that this A is identical to (22.6) when $\tau_i = 0$ for all i. Now, suppose that a_i and $(1-\tau_i)$ follow a bivariate lognormal distribution. In particular, $(\log(a_i), \log(1-\tau_i)) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\nu_a - \sigma_a^2/2, \nu_\tau - \sigma_\tau^2/2)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_a^2 & \rho\sigma_a\sigma_\tau \\ \rho\sigma_a\sigma_\tau & \sigma_\tau^2 \end{bmatrix}.$$

Plugging this formulation into (22.10), we obtain[9]

$$A = \exp\left(\nu_a + \frac{\gamma}{1-\gamma} \frac{1}{2} (\sigma_a^2 - \sigma_\tau^2) \right). \tag{22.11}$$

[9]See Appendix 22.A.3 for derivation. Hsieh and Klenow (2009) obtain a similar expression. This case is special in that ρ does not appear in the expression for aggregate productivity. In general, the correlation between $(1-\tau_i)$ and a_i matters for the aggregate productivity, and Restuccia and Rogerson (2008) emphasize the importance of such correlation. See Hopenhayn (2014b) for related discussions.

One can easily see that this expression is identical to (22.8) when $\sigma_\tau = 0$. A large dispersion in $(1 - \tau_i)$ is detrimental to aggregate productivity. This result is because, when $(1 - \tau_i)$ is dispersed, some productive firms do not expand, because $(1 - \tau_i)a_i$ is low even when a_i is large. At the same time, some unproductive firms employ a large amount of input because $(1 - \tau_i)a_i$ is large even though a_i is small. Note that, in this setting, ν_τ does not influence aggregate productivity because it does not distort the allocation of input across firms, and the total supply of inputs is fixed.

Another subset of literature examines various specific policies and institutions that cause misallocation. Examples include size-specific taxes and regulations, entry regulations, and regulations governing hiring and firing.

22.5 Firm heterogeneity in general equilibrium

When firms are forward-looking, frictions for reallocation have further effects through the firms' behavior. Hopenhayn and Rogerson (1993) highlight this mechanism in the context of firing taxes and quantify the outcome in a general equilibrium framework. In the following, we introduce the Hopenhayn and Rogerson (1993) framework with slightly different notations. In addition to the experiments on firing taxes (replicating the Hopenhayn and Rogerson, 1993 exercises), this section conducts experiments with entry barriers.[10]

22.5.1 Setup

There is a continuum of firms in the economy. We focus on the steady state, where prices and aggregate quantities (employment, output, and the number of firms) remain constant over time. In this section, we omit the firm's index i when there is no risk of confusion. Each firm uses only labor ℓ_t as an input. Firms behave competitively and maximize their profit with wage w_t. The firm's production function is $y_t = a_t \ell_t^\gamma$, where a_t is (exogenous) idiosyncratic productivity. In addition to wages, firms must pay c_f units of goods as the fixed operation cost every period. The firing taxes imposed by the government take the form of $\tau \max(0, \ell_{t-1} - \ell_t)$, where $\tau > 0$ is the firing tax for dismissing one worker. The government transfers all firing taxes back to the representative consumer. Therefore, the firm's flow profit is

$$\pi(\ell_{t-1}, \ell_t, a_t) = a_t \ell_t^\gamma - w_t \ell_t - c_f - \tau \max(0, \ell_{t-1} - \ell_t).$$

Note that the output price is normalized to 1, and the only endogenous price in each period is w_t.

The timing for the firms within a period is as follows. At the beginning of each period, the incumbent firm from the last period decides whether to exit. If the firm exits, it pays the firing cost $\tau \ell_{t-1}$. If it stays, it receives the current period value of a_t from the stochastic process

$$\log(a_t) = \alpha + \rho \log(a_{t-1}) + \varepsilon_t,$$

[10] The analysis of entry barriers is not in Hopenhayn and Rogerson (1993) but is subsequently conducted by, for example, Moscoso Boedo and Mukoyama (2012).

where α and $\rho \in [0, 1)$ are parameters and $\varepsilon_t \sim N(0, \sigma^2)$. After observing a_t, the firm decides on its employment and then produces.

Note that, unlike the model in Section 22.1, the firm's employment decision is dynamic in the presence of a positive firing tax. When the firm decides to hire a worker, it foresees that it has to pay the firing cost when it wants to shed workers in the future due to a negative productivity shock. This effect makes the firm reluctant to hire a worker when it receives a positive productivity shock. The dynamic programming problem for the firm is

$$W(a, \ell_{-1}) = \max_{\ell} \pi(\ell_{-1}, \ell, a) + \beta \max(\mathbb{E}[W(a', \ell)|a], -\tau\ell), \tag{22.12}$$

where the subscript -1 represents the previous period value and prime ($'$) represents the next period value. $\beta \in (0, 1)$ is the consumer's discount factor (which is equal to the firm's discount factor) and $\mathbb{E}[\cdot|a]$ represents the expected value given a.

We assume free entry; that is, anyone can enter as long as the entry cost is paid. After the entry cost is paid, the firm draws productivity, employs workers, and produces. The entry cost is assumed to be $c_e + \kappa$, where c_e is the technological entry cost, including the investment required when entering, and κ is additional (wasteful) policy-related cost that we interpret as "entry barriers." Free entry implies

$$W^e = c_e + \kappa, \tag{22.13}$$

where W^e is the value of the entry that satisfies

$$W^e = \int (W(a, 0) + c_f) d\nu(a),$$

where $\nu(a)$ is the exogenous distribution of a for a new entrant. Note that we assume entrants do not have to pay the fixed operation cost c_f.

The representative consumer owns the firms, works, and consumes. The utility is

$$\sum_{t=0}^{\infty} \beta^t [u(C_t) - \chi L_t^s],$$

where $u(C_t)$ is increasing and concave utility from consumption C_t, $\chi > 0$ is a parameter, and L_t^s is the labor supply. In the steady state, the consumer's problem is static:

$$\max_{C, L^s} \; u(C) - \chi L^s$$

subject to

$$C \leq wL^s + \Pi + R,$$

where Π is the total profit of the firms and R is the total transfer. The first-order condition is

$$wu'(wL^s + \Pi + R) = \chi. \tag{22.14}$$

Therefore, the labor supply is a function of w and $\Pi + R$.

As Kaas (2021) points out, the competitive equilibrium of this model has a structure often referred to as "block recursive." That is, the equilibrium price (wage in this model) can be

Table 22.1: Model results with firing taxes

	$\tau = 0$	$\tau = 0.1$	$\tau = 0.2$
Wage	1.000	0.977	0.957
Total output	100	97.7	95.7
Total employment	100	98.3	97.4
Labor productivity	100	99.4	98.3
$JC(= JD)$ rate	0.28	0.25	0.21

computed without the information on the distribution of state variables across incumbent firms. To see this, note that $W(a, \ell_{-1})$ can be computed from (22.12) once the value of w is known. Thus, W^e can be computed as a function of w. The free-entry condition (22.13) can be used to pin down the equilibrium w. For a given mass of entry M, the decision rule of (22.12) can be used to compute the stationary distribution of firms across different state variables a and ℓ_{-1}. With the stationary distribution, one can compute the total labor demand L^d, the total profit Π, and the total firing tax R. All L^d, Π, and R are functions of the entry mass M. Thus, the labor supply equation (22.14) can be used to determine the level of entry mass M that is consistent with the labor market equilibrium $L^d = L^s$.

Hopenhayn and Rogerson (1993) calibrate the model with $\tau = 0$ to the U.S. economy and examine the effect of τ quantitatively. The model here is identical to theirs except that (i) some notations are different, and (ii) they normalize the wage as 1 and the market equilibrium determines the product price p, which corresponds to $1/w$ in our notation. We also set the baseline $\kappa = 0$.

The calibration procedure follows Hopenhayn and Rogerson (1993). One period is set at five years. First, the functional form of the utility function for consumption is assumed to be natural log: $u(c) = \log(c)$. Some parameters are set ex-ante. The discount factor is set at $\beta = 0.8$, corresponding to the value of 4% per year. The production function parameter γ is 0.64, corresponding to the labor share.

To set other parameters, we assume that the $(\tau, \kappa) = (0, 0)$ case corresponds to the U.S. economy and find the parameter values so that various statistics from the model-generated data match the corresponding data moments. The parameters for the productivity process are set using the property of the model that the property of the productivity shock is directly reflected in the firm's employment decision. By using the plant-level data from the U.S. manufacturing, $\alpha = 0.076$ so that the average size of the firms is 61.7 (the actual model moment is 62.4), $\rho = 0.93$ so that the autocorrelation of $\log(\ell)$ is 0.93, and $\sigma = 0.253$ so that the variance of the growth rate for ℓ is 0.53. The operation cost $c_f = 18.0$ so that the exit rate is 37% (the actual model moment is 34%). The entrant's productivity distribution ν is set so that the size distribution of young firms matches the U.S. data. The entry cost $c_e = 9.04$ so that the free-entry condition holds with $w = 1$. The disutility of working χ is set so that the steady-state labor supply L is 0.6.

22.5.2 The effects of firing taxes

Table 22.1 summarizes the steady-state outcomes of the model with $\tau = 0$, $\tau = 0.1$, and $\tau = 0.2$, keeping $\kappa = 0$. Because one period is assumed to be five years, and the period wage (earnings per worker) with $\tau = 0$ is 1, $\tau = 0.1$ corresponds to six months' salary of a worker.[11] For total output, total employment, and labor productivity, the $\tau = 0$ case is normalized to 100.[12]

There are several important points to note. First, it is not a priori obvious whether the equilibrium employment L goes up or down when τ increases. The reason is that the effect on firing (the firms fire less because of the taxes) brings L up, whereas the effect on hiring (the firms do not hire much even with a positive a shock, given that, in the future, the firm may have to fire these extra workers) brings L down. Which one dominates is a quantitative question; here, the latter effect dominates, and L decreases when τ increases to $\tau = 0.1$ and $\tau = 0.2$.

Second, the productivity Y/L declines with τ. The reason is the misallocation mentioned in Section 22.4. Because a firm with a good a shock does not expand as much as the first-best, and a firm with a bad a does not fire as many workers with the firing tax, labor is not allocated properly across firms. These incentives imply that the marginal product of labor is dispersed (in the first best, the marginal product of labor is equalized). The difference from Section 22.4 is that the misallocation stems from the firm's dynamic decisions, especially for hiring. Firing and exiting behaviors are also affected by dynamic considerations. In the general equilibrium, misallocation also affects the wage level and entry of firms.

Third, the job creation (JC) rate, defined as (22.9), decreases with τ.[13] The reallocation of labor across firms is reduced because of the reluctance to hire and fire described above. The lack of reallocation is, therefore, closely linked to productivity loss due to misallocation.

22.5.3 The effects of entry barriers

Table 22.2 describes the model outcome for $\kappa = 0$, $\kappa = 0.5$, and $\kappa = 5.0$, keeping $\tau = 0$.[14] As in Table 22.1, for total output, total employment, and labor productivity, $\kappa = 0$ case is normalized to 100.

Entry barriers have a substantial effect on the model outcome. A higher cost of entry implies that the value of firms has to be higher in equilibrium (see equation (22.13)), and

[11] Moscoso Boedo and Mukoyama (2012) computes the costs of business regulations corresponding to τ that explicitly shows up in the World Bank's Doing Business dataset. The cross-country median of τ is about eight months of annual wages, and the average of low-income countries is 1.2 times the annual wages.

[12] Although the calibration is the same as Hopenhayn and Rogerson (1993), the numbers are not exactly the same. The reason for the discrepancy is likely due to detailed differences in computation.

[13] Here, because the economy is in the steady state, the job creation rate is equal to the job destruction rate.

[14] Moscoso Boedo and Mukoyama (2012) measures the costs of entry regulations corresponding to κ in the World Bank's Doing Business dataset. The cross-country median value of κ is 3.4 times the annual wages (about 0.7 times the five-year wages), and the average of the low-income countries is 29.9 times the annual wages (corresponding to 6 times the five-year wages. Note that although κ in the current model is not in terms of annual wages, the baseline annual wage is set at 1.0, and thus the units are comparable.

Table 22.2: Model results with entry barriers

	$\tau = 0$	$\kappa = 0.5$	$\kappa = 5.0$
Wage	1.000	0.986	0.879
Total output	100	98.6	87.9
Total employment	100	99.5	96.2
Labor productivity	100	99.1	91.4
$JC(= JD)$ rate	0.28	0.28	0.28

a high value of firms implies that the equilibrium wage has to be low. A low wage affects labor productivity through three channels. First, a low wage implies that low-productivity firms are less likely to exit. This effect pushes down aggregate productivity. Second, a low exit rate means that the entry rate is also low in the steady state. Because entrants are less productive than incumbents, this effect increases aggregate productivity. Finally, the size of incumbents is larger because of lower wages, and due to decreasing returns to scale, a large scale implies lower productivity. The first and third effects push down aggregate productivity, and the second effect pushes up; the lowering force dominates quantitatively.

The outcome of this exercise also highlights heterogeneity in policy effects across different firms. Whereas entry barriers harm entry, they increase the value of incumbent firms. When considering the policies on entry, a significant conflict arises between incumbent firms and potential entrants.[15]

22.6 Alternative market arrangements

The above discussions have assumed that all markets are perfectly competitive. We have seen in earlier chapters that many macro models consider alternative market arrangements. This section introduces two alternative market arrangements with market power in the context of firm heterogeneity. There are two takeaways from this section. First, the insights on misallocation in Section 22.4 go through with minor modifications. Second, the inclusion of market power enables us to examine how firm heterogeneity interacts with other macroeconomic variables of interest, such as the aggregate level of markups.

22.6.1 Monopolistic competition

A popular alternative formulation in the macroeconomic context is monopolistic competition (Section 6.3.5). In this setting, firms produce differentiated goods, and only one firm produces each good.

A popular setting considers two types of goods, the *final good* and the *intermediate goods*. The final good is produced in a perfectly competitive sector with constant returns to scale

[15] See Mukoyama and Popov (2014) for a politico-economic analysis of policies on firm entry. Mukoyama and Popov (2014) theoretically demonstrate that the political lobbying of incumbent firms and the economic benefits from limited entry may reinforce each other, potentially generating multiple steady states.

technology:

$$Y = \left[\int y_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}. \tag{22.15}$$

where σ is the elasticity of substitution parameter. We assume $\sigma > 1$ so that the monopoly problem of each intermediate-good producer is well-defined. The aggregate Y in (22.15) can alternatively be considered as the utility by a consumer. This aggregation (22.15) is sometimes referred to as the Dixit-Stiglitz utility function.

Consider the setting in Section 22.1, except that (22.15) replaces (22.4) and that each good i is now monopolistically produced by an intermediate-good producer i. The intermediate-good producer's production structure is the same as in Section 22.1. The intermediate-good producers use the same inputs, and the input market is perfectly competitive. The aggregation of input (22.5) remains the same.

First, consider the cost-minimization problem of the final good producer:

$$\min_{\{y_i\}} \int p_i y_i di$$

subject to (22.15) for a given Y. Letting the Lagrange multiplier of the constraint be λ, the first-order condition is

$$p_i = \lambda y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}}, \tag{22.16}$$

which implies

$$\int p_i y_i di = \lambda Y$$

and thus λ represents the (minimum) cost of producing one unit of the final good. Because the final-good market is perfectly competitive, λ is also the price of the final good. Let us call this price P. From (22.16),

$$P = \left[\int p_i^{1-\sigma} di \right]^{\frac{1}{1-\sigma}}$$

holds. As in Section 22.1, normalize the final-good price to be 1. Therefore, $P = \lambda = 1$ and the inverse demand function for good i is, from (22.16),

$$p_i = y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}}. \tag{22.17}$$

Keep in mind here that Y here is a shorthand that represents the production of all other goods in (22.15).

The monopolistic producer i maximizes profit given the inverse demand function (22.17) and its production function. The problem, which corresponds to (22.1) in Section 22.1, is

$$\max_{m_i} \left(a_i m_i^{\gamma}\right)^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} a_i m_i^{\gamma} - c m_i. \tag{22.18}$$

In the Nash equilibrium, each intermediate good firm i has to solve this problem given the other firms' input choices: m_j for all $j \neq i$. The important assumption in the monopolistic competition is that each firm is small compared to the aggregate, so that it ignores the effect

of m_i (therefore y_i) on Y. Therefore, each firm takes Y as given. With a similar step as in Section 22.1, we can obtain the relationship

$$Y = \frac{c}{\gamma}\frac{\sigma}{\sigma-1}M$$

instead of (22.3), and the same expression for the aggregate production function (22.7) where A is now modified to

$$A \equiv \left(\int a_i^{\frac{1}{\frac{\sigma}{\sigma-1}-\gamma}} di\right)^{\frac{\sigma}{\sigma-1}-\gamma}$$

instead of (22.6). Note that we have $\sigma/(\sigma-1)$ instead of 1, reflecting that each firm faces another factor (in addition to the decreasing returns to scale) that limits firm size. One important difference between this formulation and the basic model in Section 22.1 is that we can now accommodate constant returns to scale (or even some increasing returns to scale).

22.6.2 Oligopoly and endogenous markups

In the model of monopolistic competition with the constant elasticity of substitution aggregation (22.15), intermediate-good producers set a constant markup. To see this, first take a look at the first-order condition of the problem (22.18):

$$\left(1-\frac{1}{\sigma}\right)\gamma a_i^{1-\frac{1}{\sigma}} m_i^{\gamma-\frac{\gamma}{\sigma}-1} Y^{\frac{1}{\sigma}} = c. \tag{22.19}$$

Using (22.17), $y_i = a_i m_i^{\gamma}$, and the fact that the marginal cost $\mathcal{M} \equiv \partial(cm_i)/\partial y_i$ can be expressed as

$$\mathcal{M} = \frac{c m_i^{1-\gamma}}{\gamma a_i}, \tag{22.20}$$

(22.19) can be rewritten as

$$p_i = \frac{\sigma}{\sigma-1}\mathcal{M}. \tag{22.21}$$

Therefore, $\sigma/(\sigma-1)$ is the markup and is constant as long as σ is constant.

In many contexts, this constant markup property is a convenient model feature. However, this feature also imposes some limitations: the model cannot be used to analyze the endogenous changes in markups when the economic environment or policies change. The question of markup determination is particularly relevant in the recent U.S. economy. As mentioned in Section 22.3, De Loecker et al. (2020) observe that the level of markups has increased since the 1980s in the U.S. economy. There are three different paths explored by researchers: (i) departing from the monopolistic competition assumption; (ii) departing from the CES assumption; and (iii) considering the endogenous difference in productivity across firms. This section provides a brief introduction to the first approach, which is based on the formulation presented in Atkeson and Burstein (2008).

Consider the same setting as in Section 22.6.1, but each intermediate good itself is the combination of several products. Following Atkeson and Burstein (2008), call the collection of J firms that produce inputs for the intermediate good i as the *sector* i. Within each sector,

let us index each firm by j and call a particular firm's product a *brand*. The production of intermediate good i is dictated by the function

$$y_i = \left[\sum_{j=1}^{J} q_{ij}^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}. \tag{22.22}$$

We assume that $\eta < \infty$, that is, brands are imperfect substitutes. We also assume that $\eta > \sigma > 1$, that is, the brands are more substitutable within the sector than across sectors. The final-good production function is (22.15). We keep assuming that each intermediate good is small compared to the entire economy so that each firm does not consider the influence of its production decision on the final good production Y (or the general price level). However, it is sufficiently large within each sector so that it is aware of the influence on y_i (and the price of intermediate good i).

In this setting, the final good producer has to solve two layers of the cost-minimization problem: (i) find the best combination of y_i for a given Y, and (ii) find the best combination of q_{ij} for a given y_i. The first cost-minimization problem is identical to the one in Section 22.6.1. The inverse demand function of y_i is given by (22.17). The second-stage cost-minimization problem can be solved similarly, and the result is

$$\frac{\hat{p}_{ij}}{p_i} = q_{ij}^{-\frac{1}{\eta}} y_i^{\frac{1}{\eta}}, \tag{22.23}$$

where $\hat{p}_{ij}$ is the price of the brand j in sector i and p_i is now the price of the sector-i good:

$$p_i = \left[\sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta} \right]^{\frac{1}{1-\eta}}.$$

Note that (22.22) and (22.23) imply

$$p_i y_i = \sum_{j=1}^{J} \hat{p}_{ij} q_{ij}. \tag{22.24}$$

which is the consequence of (22.22) being a constant returns to scale function. Let the production function for firm j in sector i be

$$q_{ij} = a_{ij} m_{ij}^{\gamma}, \tag{22.25}$$

where m_{ij} is the "combined input" as before. The firm maximizes profit $\hat{p}_{ij} q_{ij} - c m_{ij}$. Using the inverse demand function, the problem the firm solves is

$$\max_{q_{ij}, m_{ij}} \; q_{ij}^{-\frac{1}{\eta}} y_i^{\frac{1}{\eta}} y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} q_{ij} - c m_{ij}.$$

subject to (22.22) (with q_{ij} for the other firms as given) and (22.25). As in the monopolistic competition case, the firm takes Y as given. From the first-order condition, noting the relationship (22.17) and (22.23), the pricing rule can be derived as

$$\hat{p}_{ij} = \frac{\varepsilon(s_{ij})}{\varepsilon(s_{ij}) - 1} \mathcal{M}, \tag{22.26}$$

where $\mathcal{M}$ is the marginal cost (analogous to (22.20))

$$\mathcal{M} = \frac{cm_{ij}^{1-\gamma}}{\gamma a_{ij}}$$

and

$$\varepsilon(s_{ij}) = \left[\frac{1}{\eta}(1 - s_{ij}) + \frac{1}{\sigma}s_{ij}\right]^{-1}. \tag{22.27}$$

Here, s_{ij} is

$$s_{ij} = \frac{\hat{p}_{ij}q_{ij}}{p_i y_i} = \frac{\hat{p}_{ij}q_{ij}}{\sum_{h=1}^{J}\hat{p}_{ih}q_{ih}}, \tag{22.28}$$

where the second equation utilizes the relationship (22.24). Thus, s_{ij} is the sales share of the firm j in sector i. Because $s_{ij} \in [0, 1]$ and $\sigma < \eta$, $\varepsilon(s_{ij})$ takes the value between σ and η. In particular, $\varepsilon(s_{ij})$ is η when $s_{ij} = 0$, monotonically decreases with s_{ij}, and reaches $\varepsilon(s_{ij}) = \sigma$ when $s_{ij} = 1$.

Comparing (22.21) and (22.26), the difference is that σ is replaced by $\varepsilon(s_{ij})$. In the current model, the markup of firm j can vary depending on the sales share. When firms are symmetric,

$$\varepsilon(s_{ij}) = \left[\frac{1}{\eta}\frac{J-1}{J} + \frac{1}{\sigma}\frac{1}{J}\right]^{-1}.$$

It is intuitive that the markup is the highest with $\varepsilon(s_{ij}) = \sigma$ when $J = 1$ (monopoly within the sector), which is exactly the monopolistic competition case (22.21). The markup decreases as J increases and $\varepsilon(s_{ij}) \to \eta$ as $J \to \infty$. This framework allows the monopoly power (represented by the number of firms J) to be linked to the markup. When firms are not symmetric (for example, the values of a_{ij} are different across firms), firm heterogeneity can feed into heterogeneity in markups.

In the above derivation, we made the Cournot competition assumption: each producer chooses its quantity given the quantities of the other producers in the same sector. Alternatively, we can make the Bertrand competition assumption: each producer chooses its price given the prices of the other producers in the same sector. It can be shown that the formula (22.26) still holds, with different $\varepsilon(s_{ij})$:

$$\varepsilon(s_{ij}) = \eta(1 - s_{ij}) + \sigma s_{ij}$$

instead of (22.27), where s_{ij} is still defined by (22.28). The intuition is similar to the Cournot competition case. See Appendix 22.A.4 for the details of the derivation.

22.7 Business cycles and heterogeneous firms

As we saw in Section 22.3, many statistics on firm behavior, such as job creation and destruction rates, as well as entry and exit rates, exhibit clear cyclical patterns. It is natural, therefore, to think of the causes and consequences of such cyclicality in firm dynamics.

Note that in the model of Section 22.5, firms face idiosyncratic shocks, but the aggregate economy is stationary. The basic logic is simple: firm-level shocks are smoothed out by being

summed up across firms to create GDP. To illustrate, suppose that the GDP Y_t is the sum of firm-level output y_{it}, $i = 0, 1, ..., N$.[16] The growth rate of y_{it} is identically and independently distributed with mean 0 and variance σ^2. That is,

$$\frac{y_{i,t+1} - y_{it}}{y_{it}} = \sigma\varepsilon_{i,t+1},$$

where $\varepsilon_{i,t+1}$ is a random variable with mean zero and variance one. Then, the growth rate of Y_t is

$$\frac{Y_{t+1} - Y_t}{Y_t} = \frac{1}{Y_t}\sum_{i=1}^{N} \Delta y_{i,t+1} = \sum_{i=1}^{N} \frac{y_{it}}{Y_t}\sigma\varepsilon_{i,t+1}.$$

Thus, the standard deviation of GDP growth rate, σ_Y, is

$$\sigma_Y = \sigma\sqrt{\sum_{i=1}^{N}\left(\frac{y_{it}}{Y_t}\right)^2}. \tag{22.29}$$

When the firms are of equal size, that is, $y_{it}/Y_t = 1/N$, $\sigma_Y = \sigma/\sqrt{N}$. When there are one million firms in the economy (there are over 5 million firms in the U.S. data), $1/\sqrt{N} = 0.001$. A typical standard deviation of the firm-level volatility is between 10% and 20%;[17] thus, the effect of idiosyncratic shocks on the aggregate volatility is about two orders of magnitude smaller than the GDP volatility in the U.S. data. In other words, it is negligible compared to the actual business cycle fluctuations.

22.7.1 Aggregate shocks and firm dynamics

A strand of literature takes the law of large numbers as given and adds aggregate shocks in analyzing aggregate fluctuations. In that case, the model in Section 22.5 can be modified to have a production function for the firm i:

$$y_{it} = z_t s_{it} \ell_{it}^{\gamma}.$$

Here, the variable z_t is added. As in the standard real business cycle model, z_t can be interpreted as the aggregate productivity shock. The model can then be calibrated and computed. As discussed earlier, the Hopenhayn and Rogerson (1993) model has a *block-recursive* structure. Therefore, computing this type of model is often substantially easier than computing standard heterogeneous-agent models.

It is known that the modified Hopenhayn and Rogerson (1993) model performs well in replicating the aggregate fluctuations in statistics such as job creation and destruction rates, as well as entry and exit rates. Some other firm-level statistics are difficult to replicate by a simple modification of the Hopenhayn and Rogerson (1993) model.[18]

[16] This illustration is based on the exposition of Gabaix (2011).
[17] Gabaix (2011) estimates it to be 12% in the U.S, data.
[18] See Lee and Mukoyama (2018) for a detailed analysis.

22.7.2 Can idiosyncratic shocks generate aggregate fluctuations?

Despite the law of large numbers result, some researchers believe that micro-level shocks can have an important role in generating aggregate fluctuations. The calculation at the beginning of this section made two assumptions: (i) the distribution of idiosyncratic shocks has a finite variance, and (ii) there are no input-output networks. This subsection introduces the analysis of the economic environment where one of these two assumptions does not hold.

This research agenda is attractive because, since the outset of the real business cycle research agenda, fluctuations in aggregate shocks have often been criticized for being a "black box." If we know that the cycles stem from idiosyncratic shocks, one can imagine that an effective stabilization policy would target the firms whose idiosyncratic shocks matter for the aggregate fluctuations.

Hulten's theorem

Before discussing how micro shocks can matter for macroeconomic fluctuations, it is useful to introduce a simple theorem by Hulten (1978). Let us consider a setting where there are N different sectors, indexed by i. Here, we use the terminology "sectors" because we would like to think of a competitive equilibrium. Using "firms" would yield the same result as long as firms behave as price-takers. Sector i's production is y_i. The production function is

$$y_i = a_i F(k_i, \ell_i, x_{i1}, x_{i2}, ..., x_{iN}),$$

where a_i is the TFP, x_{ij} is the sector-j product used in sector i, k_i is capital used in sector i, and ℓ_i is labor used in sector i. Note that the total sales $\sum_i p_i y_i$ is different from the total value added (i.e., the GDP), which is equal to $\sum_i p_i c_i$, because some of the output is used as intermediate goods. Let $Y = \sum_i p_i c_i$.

Hulten's (1978) theorem states that the output effect of

$$\frac{dY}{Y} = \sum_i D_i \frac{da_i}{a_i},$$

where D_i is the weight called the Domar weight:

$$D_i = \frac{p_i y_i}{\sum_i p_i c_i}. \tag{22.30}$$

D_i's denominator is the total value added, whereas the numerator is the sales of sector i. The proof of the theorem can be found in Appendix 22.A.5.

Two pieces of intuition are key to understanding Hulten's theorem. First, why does only a_i matter and not anything about inputs? The intuition is the envelope theorem. Because the economy achieves the first best, the input allocation is already optimized. Therefore, to the first-order approximation, the adjustment of inputs due to shocks to a_i does not have an impact on welfare. With the homothetic utility, the welfare result can be mapped to GDP. Second, why is the numerator of the weight measured as sales? This question seems natural, especially because the denominator is in value added, which implies that the Domar weight does not necessarily sum up to 1. The intuition is that, when there are

input-output networks, the improvement in the TFP in a downstream firm also raises the value of intermediate inputs.

To see the second point more clearly, consider the following simple example. Suppose that there are two sectors, sectors 1 and 2. Sector 1 produces the consumption good, whose price is normalized to 1. The production function is $y_1 = a_1 x_1^{1-\gamma} \ell^{\gamma}$, where y_1 is the output, a_1 is the TFP, x_1 is the intermediate input, and ℓ is the labor input. The parameter $\gamma \in (0, 1)$. Sector 2 produces the intermediate good using capital: $y_2 = a_2 k$. The capital supply is fixed at K and the labor supply is fixed at L. In the competitive equilibrium, because $Y = a_1 (a_2 K)^{1-\gamma} L^{\gamma}$,

$$\frac{dY}{Y} = \frac{da_1}{a_1} + (1-\gamma)\frac{da_2}{a_2} \tag{22.31}$$

holds. Note that the weights in front of both TFP growths do not sum up to 1. To confirm Hulten's Theorem, let us compute the Domar weight of each sector. Let the price of the intermediate good be p. In the competitive equilibrium, $p = (1-\gamma) a_1 x_1^{-\gamma} L^{\gamma}$ holds, where $x_1 = a_2 K$. Thus, the value added of sector 2 is

$$V_2 = (1-\gamma) a_1 (a_2 k)^{1-\gamma} L^{\gamma}.$$

The value added of sector 1 is

$$V_1 = Y - p x_1 = \gamma a_1 (a_2 k)^{1-\gamma} L^{\gamma}.$$

Thus, the Domar weight of sector 2 (because the sales are equal to the value added in sector 2) is $V_2/Y = (1-\gamma)$. The Domar weight of sector 1 (because the sales are Y) is $Y/Y = 1$. Both correspond to the coefficients on the right-hand side of (22.31), confirming Hulten's theorem. Note that if we consider the value added instead of sales in sector 1, the coefficient is computed as $V_1/Y = \gamma$.

Large firms

As we discussed in Section 22.3, there are many large firms in the U.S. economy. In fact, it is known that the U.S. firm size distribution is "fat-tailed": it can be closely approximated by a Pareto distribution at the right tail. Figure 22.9 reproduces the data plot of Figure 2A in Carvalho and Grassi (2019). It plots the size of a firm, measured by employment, against the firm size "percentile" (the fraction of firms whose size is larger than the value on the x-axis). The triangles use the same information as in Figure 22.1.[19] Because the publicly available BDS table does not contain finer information on large firms, Carvalho and Grassi (2019) utilizes the Compustat data, which includes only publicly traded firms (circles). To the extent that large privately-held firms are rare, the red circles approximate the distribution of large firms in the U.S. economy. One can see that the plot with log-log axes is close to a straight line, indicating that the distribution is well-approximated by the Pareto distribution.

Gabaix (2011) argues that when the distribution of firm size is fat-tailed, that is, there is a considerable presence of large firms (as in the Pareto distribution), the formula (22.29)

[19] The time period for Figure 22.9 is 1977–2012, whereas the time period for Figure 22.1 is 2019. Figure 22.9 plots in finer categories than Figure 22.1 does.

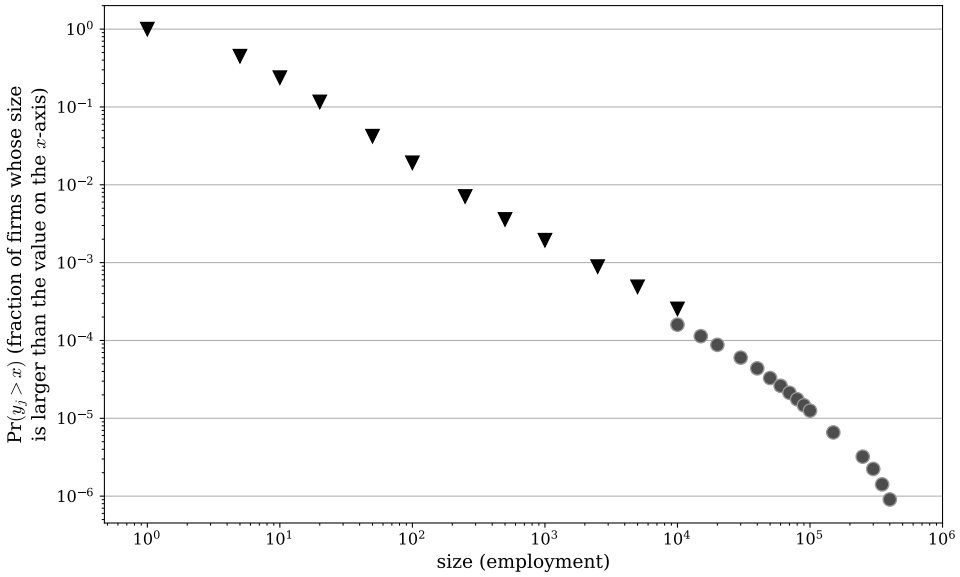


Figure 22.9: Distribution of firm size in log-log axes.

Source: Business Dynamics Statistics and Compustat. Reproduced from Carvalho and Grassi (2019).

implies a significant impact of σ on σ_Y. In particular, he considers the case where the firm size distribution is Pareto

$$\Pr[y_i > x] = \chi x^{-\zeta}, \tag{22.32}$$

where χ and ζ are constants. The variable y_i here is employment, and $\Pr[y_i > x]$ is the fraction of firms whose sizes are larger than x. It is known that the U.S. firm size distribution has ζ close to 1. (Because (22.32) implies $\log(\Pr[y_i > x]) = -\zeta \log(x) + \log(\chi)$, this can be seen from the slope of Figure 22.9 being close to -1.) Therefore, the empirically relevant case is $\zeta = 1$. In this case, he derives

$$\sigma_Y \sim \frac{v_\zeta}{\log(N)}\sigma$$

as $N \to \infty$, that is, $\sigma_Y \log(N)$ converges to $v_\zeta \sigma$ in distribution, where v_ζ is a random variable that does not depend on N or σ. This formula implies that the GDP volatility is proportional to $1/\log(N)$, rather than $1/\sqrt{N}$. The function $1/\log(N)$ declines slower than $1/\sqrt{N}$ as N becomes large (for example, $1/\log(1,000,000) \approx 0.072$ whereas $1/\sqrt{1,000,000} = 0.001$). Gabaix (2011) calculates that the coefficient on σ in the above formula would be approximately 0.12 instead of 0.001. The effect of idiosyncratic shocks is, therefore, two magnitudes larger than the identical-firm case; thus, the volatility induced by the idiosyncratic shock can have an effect of a similar magnitude to observed business cycles. In a recent work, Carvalho and Grassi (2019) build a quantitative business cycle model similar to the one in Section 22.5 and analyze the business cycle dynamics driven by idiosyncratic shocks to large firms.

Effects of production networks

An additional important factor that can magnify the effect of idiosyncratic shocks on the aggregate economy is that the Domar weight (22.30) divides the firm sales by the aggregate value added. Many large firms, such as Walmart, Amazon, and GM, have significantly larger sales than their value added, magnifying their contribution to the aggregate fluctuations. These firms are at the downstream of the production network, and the impact of their productivity shocks on the aggregate GDP is larger than their value added.

Some researchers believe that there are further important implications of the production network. Note that Hulten's theorem builds on two assumptions: (i) the economy is efficient, and (ii) the first-order approximation is sufficiently accurate. When these two assumptions are not appropriate, network structures can play a role in considering the aggregate effects of idiosyncratic shocks. For example, in a recent paper, Baqaee and Farhi (2019) argue that there are situations where the second-order effect is quantitatively important. In another paper, Baqaee and Farhi (2020) consider economies with distortions, and there the network structure also plays a role.

22.8 Endogenous productivity

Given the importance of the idiosyncratic shocks in generating heterogeneity across firms and their dynamics, it is natural to wonder what factors influence idiosyncratic productivity. One natural framework that can be used to analyze this issue is the models of endogenous productivity change, introduced in the economic growth chapter (Chapter 13). Klette and Kortum (2004) introduced a framework that can generate the entry, exit, expansion, and contraction of firms due to endogenous innovation. Below, we explain the Klette and Kortum (2004) model with discrete time.[20]

Consider an economy with a continuum of products on the unit interval $[0, 1]$. We will focus on the balanced-growth path of the economy, where all aggregate variables grow at the same rate. The representative consumer's utility function is

$$\sum_{t=0}^{\infty} \beta^t \log(C_t),$$

where $\beta \in (0, 1)$ is the discount factor and C_t is defined as

$$C_t = \exp\left(\int_0^1 \log\left(\sum_{k=-1}^{J_t(j)} q_t(j,k) c_t(j,k) \right) dj \right). \tag{22.33}$$

Here, $j \in [0, 1]$ is the index of the product. The aggregation with natural log implies that the elasticity of substitution across goods is 1. The index k is an integer that runs from -1 to $J_t(j)$. The value of k represents the generation of the product: a newer generation (i.e., a larger k) product is of higher quality. $J_t(j)$ is the cutting-edge (state-of-the-art) generation of good j. Generation k of product j (call it product (j, k) for short) contributes to consumption

[20] Ates and Saffie (2021) also develop a discrete-time version of the Klette and Kortum (2004) model.

in the form of $q_t(j,k)c_t(j,k)$. Here, $q_t(j,k)$ is the *quality* of the product (j,k) and $c_t(j,k)$ is the *quantity* of the product (j,k). The fact that the aggregation is additive across different generations implies that different generations are perfect substitutes. If generation k' has twice the quality of generation k, that is, $q_t(j,k')/q_t(j,k) = 2$, consuming one unit of product (j,k') is equivalent to consuming two units of product (j,k).

The consumer faces two layers of problems: intratemporal and intertemporal. The intratemporal problem is to determine how to allocate expenditure across different goods at each point in time. The intertemporal problem is to decide how much to spend across time.

Let us start by thinking about the intratemporal problem. Let E_t be the expenditure at period t and $p_t(j,k)$ be the price of product (j,k). First note that within product j, it is optimal to purchase only the generation whose "quality-adjusted price" $p_t(j,k)/q_t(j,k)$ is the lowest. Therefore, let us only consider such a generation k for each j. Then, the intratemporal problem is

$$\max_{c_t(j,k)} \int_0^1 \log\left(q_t(j,k)c_t(j,k)\right) dj$$

subject to

$$\int_0^1 p_t(j,k)c_t(j,k)dj \leq E_t. \tag{22.34}$$

The solution is

$$c_t(j,k) = \frac{E_t}{p_t(j,k)}. \tag{22.35}$$

Given the solution to the intratemporal problem, the consumption of each period can be rewritten as

$$C_t = E_t \exp\left(\int_0^1 [\log(q_t(j,k)) - \log(p_t(j,k))]dj\right).$$

This relationship can be rewritten as

$$P_t C_t = E_t,$$

where the price index for consumption is

$$P_t \equiv \exp\left(\int_0^1 [\log(p_t(j,k)) - \log(q_t(j,k))]dj\right).$$

As in Section 22.6.1, normalize $P_t = 1$. This normalization implies

$$\int_0^1 \log(p_t(j,k))dj = \int_0^1 \log(q_t(j,k))dj \tag{22.36}$$

at any t.

The intertemporal problem is now

$$\max_{C_t} \sum_{t=0}^{\infty} \beta^t \log(C_t)$$

subject to

$$\sum_{t=0}^{\infty}\left(\frac{1}{1+r}\right)^{t} C_t \leq \mathcal{A}_0,$$

where $\mathcal{A}_0$ is the present discounted value of all future labor and asset incomes. The asset income in this economy comes from the claim to the profit of the firms. Here, we are imposing that the interest rate $r > 0$ is constant, as we will focus on the balanced growth path. The optimization results in the Euler equation:

$$\frac{1}{C_t} = \beta(1+r)\frac{1}{C_{t+1}}.$$

Along the balanced-growth path, C_{t+1}/C_t grows at a constant rate. Let us define $\gamma \equiv (C_{t+1} - C_t)/C_t$. Then,

$$\frac{1}{1+r} = \frac{\beta}{1+\gamma}. \tag{22.37}$$

Each product (j,k) is produced by one firm. One firm can own several product lines. A *firm* here is indeed defined as a collection of product lines. A small firm owns only a few product lines, and a large firm owns many product lines. Although firms are heterogeneous in this dimension, the analysis of the firm decision in the Klette and Kortum (2004) model is relatively simple because the model has a structure that allows each of the firm's product lines to make decisions independently. In the following, we exploit this property and analyze the firm's decisions at the product line level.

First, consider the production decision. Producing one unit of a product takes one unit of labor.[21] Thus, the unit production cost is w_t. Given that the production cost is the same, the cutting-edge producer (the "leader") has an advantage over other producers (with the older generation of product j). Because the demand elasticity is 1, the optimal pricing is to set the price as high as possible. Here, the cutting-edge producer cannot increase the price in an unlimited manner. Once the price is set sufficiently high, the $J_t(j)-1$ generation producer can enter profitably. The highest price the cutting-edge producer can charge is

$$p_t(j, J_t(j)) = \lambda w_t, \tag{22.38}$$

where $\lambda > 1$ represents the technology step $q_t(j,k+1)/q_t(j,k)$. Here λ coincides with the markup rate. This pricing behavior is called *limit pricing*.

Given the price, the period profit from one product line for the leader is

$$\pi_t \equiv (p_t(j, J_t(j)) - w_t)\frac{C_t}{p_t(j, J_t(j))} = \left(1 - \frac{1}{\lambda}\right) C_t. \tag{22.39}$$

From (22.36) and (22.38),

$$\log(\lambda w_0) = \int_0^1 \log(q_0(j, J_0(j))dj.$$

[21] As we saw in Section 22.6.1, the production scale remains finite even with constant returns to scale because of the monopoly power.

By normalizing $q_0(j, J_0(j)) = \lambda$ for all j, this relationship implies $w_0 = 1$.

Second, consider the innovation decision. We assume that innovation is also conducted in each product line. In each product line, the firm decides the intensity of innovation η with the required labor input $R(\eta)$; therefore, the innovation cost is $w_t R(\eta)$, where $R(\cdot)$ is an increasing and convex function. The intensity η represents the probability that the firm gains another product line. When innovation is successful, in addition to the current product line, the firm can start producing using another product line. This newly-added product is λ times better than the current cutting-edge product. This newly-added product is randomly chosen from $[0, 1]$. Each firm is infinitesimally small compared to the entire economy; thus, the probability of innovating over its own product is zero, and the new innovation always takes the market from another firm. Because other firms also innovate, each firm can also lose its market for the product line. Let μ be the probability that other firms innovate and take over the current product line. Along the balanced-growth path, μ is constant over time.

Denoting the value of a leader product line by V_t, the Bellman equation for the firm is

$$V_t = \max_{\eta} \ \pi_t - w_t c(\eta) + \frac{1}{1+r}(1 + \eta - \mu)V_{t+1}. \tag{22.40}$$

The final term is the expected future value of the product line. There are four possible scenarios for the current leader of the product j: (i) it innovates and is not taken over by another firm; (ii) it fails to innovate and is taken over by another firm; (iii) both occur; and (iv) neither occurs. The probability of (i) is $\eta(1 - \mu)$ and the future value is $2V_{t+1}$. The probability of (ii) is $\mu(1 - \eta)$ and the future value is 0. The probability of (iii) is $\mu\eta$ and the future value is V_{t+1}. The probability of (iv) is $(1 - \mu)(1 - \eta)$ and the future value is V_{t+1}. Therefore, the expected future value is computed as $(1 + \eta - \mu)V_{t+1}$, which can be seen in the final term.

Along the balanced-growth path, V_t, π_t, and w_t all grow at the common rate $(1 + \gamma)$. Dividing both sides of (22.40) by $(1 + \gamma)^t$ and using (22.37), (22.39), and $w_0 = 1$,

$$v = \max_{\eta} \ \left(1 - \frac{1}{\lambda}\right) C_0 - R(\eta) + \beta(1 + \eta - \mu)v, \tag{22.41}$$

where $v \equiv V_t/(1 + \gamma)^t$. The first-order condition is

$$R'(\eta) = \beta v. \tag{22.42}$$

There are many potential entrants in the economy. As in Section 22.5, we assume free entry. Entrant (with probability 1) can hire c_e units of labor and enter. The free-entry condition is

$$V_t = w_t c_e.$$

Normalizing,

$$v = c_e. \tag{22.43}$$

Let the entry rate (the amount of entry at each period) be ν. We assume that each product line receives only (up to) one innovation per period. Thus, the fraction of the product lines that receive innovation is μ. Because the innovation is done by either entrants or incumbents,

$$\mu = \eta + \nu \tag{22.44}$$

holds.

There are three types of labor demand: (i) production, (ii) innovation by incumbents, and (iii) entry. From (22.35) and $E_t = C_t$, production at time 0 is $c_0(j,k) = C_0/\lambda$ and thus the labor demand is C_0/λ. For innovation by incumbents, $R(\eta)$ units of labor are used. For entry, ν units are demanded. The labor supply is fixed at L. Thus, the labor-market equilibrium condition (using (22.44)) is

$$\frac{C_0}{\lambda} + R(\eta) + \nu = L. \tag{22.45}$$

In sum, the general equilibrium of the model solves four unknowns (v, η, C_0, and μ) with four equations:

$$v = \left(1 - \frac{1}{\lambda}\right) C_0 - R(\eta) + \beta(1 + \eta - \mu)v,$$

which is from (22.41), the first-order condition (22.42), the free-entry condition (22.43), and the labor-market equilibrium condition (22.45).

Finally, we can calculate the economy's growth rate, γ. In this economy, the consumption C_t in (22.33) is equal to E_t, which is aggregate expenditure (see (22.34)). Along the balanced-growth path, the growth rate of E_t is also equal to the growth rate of w_t. From (22.36) and (22.38),

$$\begin{aligned}
(\gamma \approx) \log(w_{t+1}) - \log(w_t) &= \int_0^1 \log(p_{t+1}(j, J_{t+1}(j))) - \log(p_t(j, J_t(j)))dt \\
&= \int_0^1 \log(q_{t+1}(j, J_{t+1}(j))) - \log(q_t(j, J_t(j)))dt \\
&= \int_0^1 (\log(\lambda^{J_{t+1}(j)}) - \log(\lambda^{J_t(j)}))dt \\
&= \mathbb{E}[\log(\lambda^{J_{t+1}})] - \mathbb{E}[\log(\lambda^{J_t})] \\
&= \mathbb{E}[J_{t+1}]\log(\lambda) - \mathbb{E}[J_t]\log(\lambda) \\
&= \mu(t+1)\log(\lambda) - \mu t \log(\lambda) \\
&= \mu \log(\lambda).
\end{aligned}$$

The first equality follows from (22.38), the second is from (22.36), and the third is from the definition that $J_t(j)$ is the cutting-edge generation at industry j. In the fourth equality, we utilize the law of large numbers. Because each industry is subject to the i.i.d. shock and there is a continuum of industries, we can replace the cross-sectional average with the expected value when we interpret J_t as a random variable. The next inequality uses the fact that J_t is viewed as a sum of Bernoulli trials with winning probability μ (i.e., every period, the probability that a product *receives* an innovation is μ). The growth rate of the economy depends on μ, which is driven by the innovation intensity by incumbents (η) and by entrants (ν), and the innovation step λ.

The major strength of the Klette and Kortum (2004) model over traditional endogenous growth models is that the definition of a firm is clear, allowing for the analysis of the dynamics of firms and the firm-size distribution. For the current discrete-time model, the analysis of firm-size distribution is somewhat more complex than the original Klette-Kortum model (which is formulated in continuous time), because in a discrete-time formulation, many events

can happen to a firm in the same period. The details are described in Appendix 22.A.6. Although the firm-size distribution is not analytically straightforward in the discrete-time version, it is straightforward to compute it on a computer. The advantage of this model over models with exogenous idiosyncratic shocks, such as the one in Section 22.5, is that it can analyze how policies and changes to the economic environment affect the productivity process itself.[22]

One simple statistic that we can compute is the average growth rate of the firm size. First, note that because each product line produces the same quantity and employs the same number of workers, the firm size distribution coincides with the distribution of product lines across firms. As can be seen in the discussion of (22.41), the average number of product lines in the next period per each line this period is $(1+\eta-\mu)$. Therefore, the average growth rate of the firm size is $\eta-\mu$. From (22.44), $\eta-\mu=-\nu<0$. Thus, the average growth rate of a firm is negative, and a large firm almost always contracts (due to the law of large numbers). The property that a large and small firm have a common average growth rate is called Gibrat's Law, although it is usually stated in the context of the positive average growth rate.

There are several counterfactual predictions in the Klette and Kortum (2004) model. Recent literature has made progress in modifying the model to replicate the salient features of the data.

First, the firm with a higher-quality product does not earn a higher profit on that product line. This feature stems from two assumptions: (i) the elasticity of substitution across goods is 1, and (ii) the technology is not cumulative. For the first point, the natural log specification implies that the leader's revenue is the same regardless of prices and quality levels. With higher substitutability, both can matter for the size and profit of the firm. For the second point, any outside firm can innovate over the state-of-the-art product at the same cost as the incumbent. Suppose the model is extended so that the incumbent firm improves its own product quality in equilibrium. In that case, there is an additional reason for the size and profit difference (therefore, the idiosyncratic productivity shock in Section 22.5) across firms.

Second, the firm-size distribution does not feature a Pareto tail. Intuitively, it is challenging to create many large firms in this economy because the firm size contracts on average. One alternative example is that, instead of a negative growth rate, a large firm has a positive constant growth rate g. Suppose, in addition, all firms receive an exit shock with the probability $\delta\in(0,1)$. Consider a very large firm so that we can ignore the integer constraint of product lines. Let us start from the firms between size n and $n+\Delta$, where Δ is a small number relative to n. When the stationary density at n is $h(n)$, the mass of firms between these sizes is approximated by $h(n)\Delta$. In the next period, the surviving mass is $(1-\delta)h(n)\Delta$. After one period, size n will grow to $(1+g)n$ and size $n+\Delta$ will grow to $(1+g)(n+\Delta)$. Thus, the mass between these new sizes will be $(1+g)h((1+g)n)\Delta$. Therefore, in the stationary distribution,

$$(1+g)h((1+g)n)\Delta=(1-\delta)h(n)\Delta$$

has to hold. Guess that the distribution is Pareto: $h(n)=Fn^{-(\zeta+1)}$, where $F>0$ and $\zeta>0$ are parameters. In particular, ζ is the tail index that showed up in Section 22.7.2. The

[22] See Mukoyama and Osotimehin (2019) for an example of such policy analysis.

above equation can then be rewritten as

$$(1+g)F((1+g)n)^{-(\zeta+1)}\Delta = (1-\delta)Fn^{-(\zeta+1)}\Delta.$$

This equality holds for any n and Δ when

$$\zeta = -\frac{\log(1-\delta)}{\log(1+g)} > 0.$$

Thus, we verified that the firm dynamics with positive (and constant) growth, combined with a constant exit rate, can be consistent with the stationary distribution that is Pareto.[23] The tail index ζ is small (i.e., a thick tail) when δ is small or g is large. One question is how we can modify the Klette and Kortum (2004) model to have a positive firm growth rate at the right tail. One possibility is to break equation (22.44).[24] For example, if some innovation *creates* new products, there can be firm expansion without contributing to μ. Suppose that the new product creation among the total innovation is ξ (i.e., among the total $\eta+\nu$ innovation, ξ creates new products, and $\mu = \eta+\nu-\xi$ replaces existing products). Then, the average growth rate of a firm, which is still $\eta-\mu$, is now equal to $\xi-\nu$ (instead of just $-\nu$). If ξ is sufficiently large, $\xi-\nu$ can be positive.

[23] See Mukoyama and Osotimehin (2019) for a similar derivation.
[24] Luttmer (2011) discusses related insights.

Chapter 23

International Macroeconomics

Giancarlo Corsetti, Luca Dedola and Simon Lloyd

23.1 Introduction

Previous chapters have examined macroeconomic dynamics within closed economies, abstracting from interactions with the rest of the world via international trade and financial flows. Yet, in reality, the global economy is deeply interconnected; no country operates in isolation. National business cycles may be impacted by spillovers from foreign shocks, while many events—like the IT revolution, the 2007-2009 Global Financial Crisis (GFC), the economic fallout of the Covid-19 pandemic, and recent geopolitical developments—are global in nature. At the same time, openness may help a country to smooth domestic shocks. All this underscores the importance of modeling open economies. In this chapter, we broaden our analysis moving from a closed economy to a financially integrated one.

We start by describing how national accounting incorporates trade in goods and financial assets between countries. We then summarize the main stylized facts of international business cycles. Grouping countries by GDP per capita, these facts lay foundations for the next two chapters: this one dedicated to the study of interactions between large advanced countries with strong institutions (i.e., those committed to repaying their debt); and the next focused on small-open economies that may default. Throughout, we show how business-cycle evidence challenges theory, highlighting numerous "puzzles" in international macroeconomics.

In this chapter, we introduce a workhorse two-country open-economy model designed to represent the dynamics of large, developed economies. We use the model to examine how TFP shocks in one country spill over to others, affect international relative prices and alter external balances. We also note that two stylized empirical facts challenge the predictions of standard models. First, net exports and the current account (i.e., external savings) are counter-cyclical; that is, external deficits become larger during booms.[1] Second, international

[1] In the conventional Mundell-Fleming model presented in undergraduate textbooks, a counter-cyclical external balance is generated with the ad hoc assumption that the 'marginal propensity to import' from income is positive, resulting in strong complementarities between foreign and domestic goods and positive cross-border transmission.

relative prices (i.e., the real exchange rate and the terms of trade) tend to appreciate when output is relatively high.

We explain how the degree of international risk sharing by households—governed by the structure of international financial markets—and the degree of substitutability between domestic and imported goods are key for reconciling these facts within the model. Following an increase in domestic productivity, a fall in the international price of home output redirects domestic and foreign demand towards goods produced in the home economy. At the same time, however, a lower price also reduces the value of any given quantity of home output—all else equal, reducing the relative income and wealth of households in the home country. We explain that model specifications in which these income fluctuations are imperfectly insured can better capture international business-cycle facts.

The core of the chapter focuses on settings with fully flexible prices. Nevertheless, the setup can be augmented to a two-country New Keynesian framework, with nominal rigidities, as we discuss briefly in our conclusion alongside other model extensions.

23.1.1 National accounting in the open economy

Recall the definition of Gross Domestic Product (GDP) from Chapter 1:

$$Y = C + I + G + NX,$$

where Y represents GDP, C is private consumption, I is investment, G is government spending and NX is net exports. NX represent the difference between a country's exports (X) and imports (M) of goods and services. They can be positive (*trade surplus*) or negative (*trade deficit*) depending on economic interactions (in goods and asset markets) with the rest of the world.

A comprehensive view of how a country interacts economically with other nations through trade, investment and financial transfers is provided by the Balance of Payments (BOP). The BOP records all international transactions. In the U.S., BOP data can be found at the Bureau of Economic Analysis, in the 'International Transactions' section.[2] BOP information for other countries is compiled by the IMF on their 'BOP information' website.[3]

The BOP has three components: the Current Account (CA), the Financial Account (FA), and the Capital Account (KA), satisfying $CA + FA + KA = 0$. The CA records transactions related to goods, services, income, and current transfers. It includes the trade balance (or net exports NX), as well income from abroad.[4] In a country like the U.S., where net factor income and net transfers from abroad are small, $CA \simeq NX$. The FA records changes in ownership of financial assets and liabilities between a country and the rest of the world.[5] The KA tracks capital transfers and the acquisition or disposal of non-produced,

[2]See https://www.bea.gov/data/intl-trade-investment/international-transactions, Table 1.1, Balance of Current Account (annual data).

[3]See https://data.imf.org/?sk=7A51304B-6426-40C0-83DD-CA473CA1FD52.

[4]The $CA = NX + NFI + NUT$, where NFI is Net Factor Income (income earned from foreign investments minus income paid to foreign investors) and NUT is Net Unilateral Transfers (gifts from foreign countries minus gifts to them).

[5]It includes Direct Investment (long-term investments in foreign businesses or assets), Portfolio Investment (investments in foreign stocks, bonds, or other financial instruments) and Other Investments (other financial transactions).

non-financial assets (e.g., patents and copyrights). It is typically smaller in magnitude compared to the other two accounts, implying that $CA \simeq -FA$.

The evolution of the BOP in the U.S. (left) and detail on the three accounts (right) are depicted in Figure 23.1.[6] The sharp decline in the external balance of the U.S. from the mid-1990s, captured in Figure 23.1 has given rise to a large literature on "global imbalances" (see, e.g., Caballero, Farhi, and Gourinchas, 2008) and a modern reconsideration of the US "exorbitant privilege" from issuing the global reserve currency (see, e.g., Farhi and Maggiori, 2017).

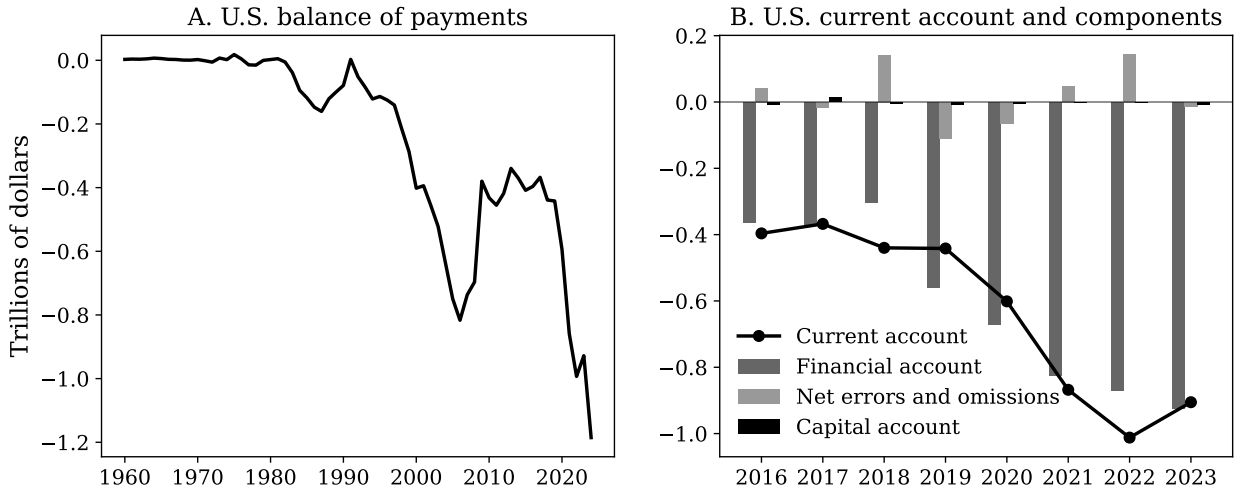


Figure 23.1: Balance of payments (left) and current, capital, and financial accounts (right) in the U.S.

Source: BEA International Transactions and IMF Data, Balance of Payments.

From national-accounting identities, we know that national saving satisfies $S = Y - G - C = I + NX$. Because $CA \simeq NX$, a current-account deficit is equivalent to a trade deficit. And since $CA \simeq -FA$, this implies that trade deficits are financed by capital inflows, as a country borrows from abroad. So, the negative current-account balance seen in the right-panel of Figure 23.1 indicates both that the U.S. has been importing more goods than it has been exporting and that it has been saving less than it has been investing domestically.

23.1.2 Cross-country differences in GDP per capita and size

Relationships between trade balances, savings and investment also shed light on broader patterns across countries, as nations with persistent trade surpluses or deficits often experience distinct trajectories of economic growth, resource allocation and external borrowing or lending. These patterns are influenced by trade specialization, where countries focus on producing goods and services in which they have a comparative advantage. This specialization shapes international trade flows, affects income levels, and contributes to the heterogeneity

[6]The plot in the left panel was downloaded frpm https://www.bea.gov/data/intl-trade-investment/international-transactions and the one in the right panel from https://data.imf.org/?sk=7a51304b-6426-40c0-83dd-ca473ca1fd52&sid=1484252556980, accessed in December 2024.

observed in GDP per capita across nations. Based on these differences, the IMF categorizes countries into three main groups: 'Advanced Economies' (AE), 'Emerging Markets' (EM), and 'Low-income and Developing Countries' (LI). In 2023, the average GDP per capita (in real 2015 US dollars) among AEs was almost \$90,000, whereas in EMs it stood at slightly above \$10,000. LI countries had an average income per capita below \$2,000. Wide disparities between these groups have been persistent over time, as Chapter 3 discussed.

Another critical dimension of heterogeneity is country size. As of 2024, the global population was approximately 7.92 billion. The two most populated nations were India (1.45 billion) and China (1.41 billion), followed by the U.S. (345 million). Europe collectively had a population of around 737 million, while Africa's population stood at approximately 1.43 billion. Countries with a relatively low GDP per capita can nonetheless have a large weight in the global economy due to their population size.

Country size has notable influence on economic interactions between regions, whether in trade of goods and services or financial-asset transactions. Larger economies can have market power, hence their macroeconomic dynamics can influence global interest rates, terms of trade and exchange rates. In contrast, smaller economies are typically price takers in international markets, responding to, rather than shaping, global economic conditions. These differences are reflected in the open-economy academic literature, which is divided into two primary strands. One branch addresses small-open economies, emphasizing their unique dynamics and constraints. These models will be explored in the next chapter. The other branch focuses on large economies and often employs models involving two countries to capture their significant mutual interactions. This approach will be the focus of this chapter where we analyze characteristics of international business cycles (i.e., fluctuations around trends) for large, advanced countries.[7]

23.2 International business-cycle facts

Open-economy business-cycle theory seeks to address fundamental questions about the behavior of aggregate economic variables within and across nations. How volatile is output? Are consumption, investment, and exports pro-cyclical or counter-cyclical? Are economic booms and recessions typically associated with trade balance deficits or surpluses? Are emerging markets different from advanced economies?

23.2.1 Macro variables

To explore business-cycle patterns globally, we extend the methodology introduced in Chapter 14, where similar statistics were computed for the U.S. We analyze a broad set of countries using country-specific data from the World Bank Development Indicators. Our panel consists of 96 countries with annual data spanning the period 1980-2023. Variable definitions,

[7] As shown by Galí and Monacelli (2005), one can derive a small-open economy version of the model in this chapter by taking limits with respect to country size, specifying one country as infinitesimally small in the global economy for which foreign quantities and prices are taken as given. Because of product specialization, however, as explained below, the country retains monopoly power on its terms of trade, i.e., it is not infinitesimal in the market for its own output.

countries included, and replication codes are provided in the book's GitHub page.[8]

To characterize the 'average' world business cycle, we compute business-cycle statistics for individual countries and aggregate them using population-weighted averages. The main statistics are presented in the second column of Table 23.1. The first column includes the values for the U.S., for comparison.[9] The table also includes the average trade-balance-to-output ratio, $tb/y = (x - m)/y$, and the average openness ratio, $(x + m)/y$, where x denote exports, m imports and y GDP (all expressed in per capita terms).

As Table 23.1 shows, national business cycles are similar to each other in numerous dimensions. In particular:

Fact 1 **Aggregate demand and its components are pro-cyclical**: consumption, investment, exports and imports are all positively correlated with output.

Fact 2 **Persistence**: output and its components—consumption, investment, exports and imports—exhibit strong persistence.

Fact 3 **Ranking of volatilities**: imports, exports and private investment are the most volatile components of GDP, followed by government spending and consumption.

Fact 4 **Trade balances and current accounts are counter-cyclical**: the trade balance, trade-balance-to-output ratio, current account and current-account-to-output ratio are negatively correlated with output. This implies that countries tend to import more than they export during expansions and run trade surpluses during recessions.

The following two facts point to significant differences between AE and EM-LI countries—setting the stage for analysis in the next chapter.

Fact 5 **Business cycles in EM and LI economies are more volatile**: the average standard deviation of output is around 2% in AEs, but the corresponding volatilities for EM and LI countries are, respectively, about 50 and 25% higher.

Fact 6 **Consumption smoothing is lower in LI economies**: private consumption, including durables, is about as volatile as output, or slightly more volatile, in AEs and EMs. This may appear counterintuitive and at odds with consumption smoothing. However, durable goods expenditures, viewed as investment in household capital, account for much of this volatility. Consumption volatility is, however, more than twice the volatility of output in LIs.

Finally, relative to the closed-economy literature, open-macro research is confronted with the issue of interdependence. Not only do countries' business cycles share similar features (i.e., for most, consumption is usually less volatile than investment), they are typically positively correlated across borders, and the correlation is especially high among advanced countries—as summarized by the following fact.

[8] It can be accessed in https://github.com/PhD-Macroeconomics/Codes-and-Data.

[9] These differ somewhat from those presented in Chapter 14 because we here use annual, rather than quarterly data, and the sample period is different to support cross-country comparison.

Table 23.1: HP-filtered Business Cycles Around the World

Statistic	US	All Economies	Advanced	Emerging Markets	Low-Income Countries
Standard Deviations					
σ_y	1.96	2.89	2.13	3.10	2.41
σ_c/σ_y	0.94	1.24	1.03	1.16	2.35
σ_g/σ_y	1.18	2.02	1.17	1.91	4.50
σ_i/σ_y	4.55	3.04	3.46	2.76	4.77
σ_x/σ_y	4.11	3.91	3.43	3.55	7.87
σ_m/σ_y	3.47	4.08	3.40	3.87	7.19
$\sigma_{tb/y}$	0.61	1.75	1.07	1.82	2.41
$\sigma_{ca/y}$	0.72	1.79	1.20	1.82	2.51
Correlations with y					
c	0.93	0.67	0.84	0.66	0.41
g	0.02	0.33	0.04	0.38	0.35
i	0.65	0.70	0.76	0.71	0.47
x	0.17	0.17	0.32	0.13	0.30
m	0.67	0.40	0.62	0.37	0.35
tb/y	-0.59	-0.27	-0.41	-0.27	-0.09
tb	-0.62	-0.28	-0.41	-0.27	-0.15
ca/y	-0.53	-0.31	-0.36	-0.32	-0.09
ca	-0.59	-0.28	-0.38	-0.28	-0.10
Auto correlations					
y	0.51	0.54	0.47	0.55	0.49
c	0.53	0.47	0.45	0.49	0.37
g/y	-0.63	-0.21	-0.61	-0.17	0.09
i	0.21	0.51	0.47	0.53	0.43
x	0.55	0.44	0.43	0.46	0.36
m	0.23	0.40	0.29	0.43	0.39
tb/y	0.62	0.41	0.50	0.41	0.20
CA/y	0.67	0.40	0.51	0.40	0.15
Means					
tb/y	-2.3	-1.0	-0.7	-0.1	-9.2
$(x+m)/y$	22.0	28.1	32.5	26.5	35.3

Notes: The variables y, c, g, i, x, m, $tb \equiv (x-m)$, and ca denote GDP, total private consumption, government spending, investment, exports, imports, the trade balance, and the current account, respectively. Variables are expressed in real per capita terms. The variables y, c, g, i, x, m are Hodrick-Prescott-filtered ($\lambda = 100$) in logs and expressed as percentage deviations from trend. The variables tb/y, g/y, and ca/y are HP-filtered in in levels. The variables tb and ca are detrended first using the long-term component of y and then HP filtered. There are 96 countries in the sample covering, approximately, the interval 1980-2023 at annual frequency. Moments are averaged across countries using long-run population weights. Countries included in Advanced Economies, Emerging Markets, and Developing Countries follow the IMF definition of these groups. The lists of countries, individual statistics, and replication material are provided the book's GitHub page: GitHub page

.

Fact 7 **Output co-moves positively across borders**: business cycles tend to be synchronized across AE countries.

Figure 23.2 illustrates this by plotting the deviations from trend in per-capita GDP for the

U.S. (solid line), Canada (dashed line) and the United Kingdom (dotted line) between 1970 and 2023. The de-trended GDP series for these countries move closely in tandem, reflecting the interconnected nature of their economies and the global transmission of economic shocks. Table 23.2 shows the average correlation of output, consumption, investment and prices between the U.S. and a subset of developed economies between 1971 and 2018, further illustrating synchronicity in international business cycles.

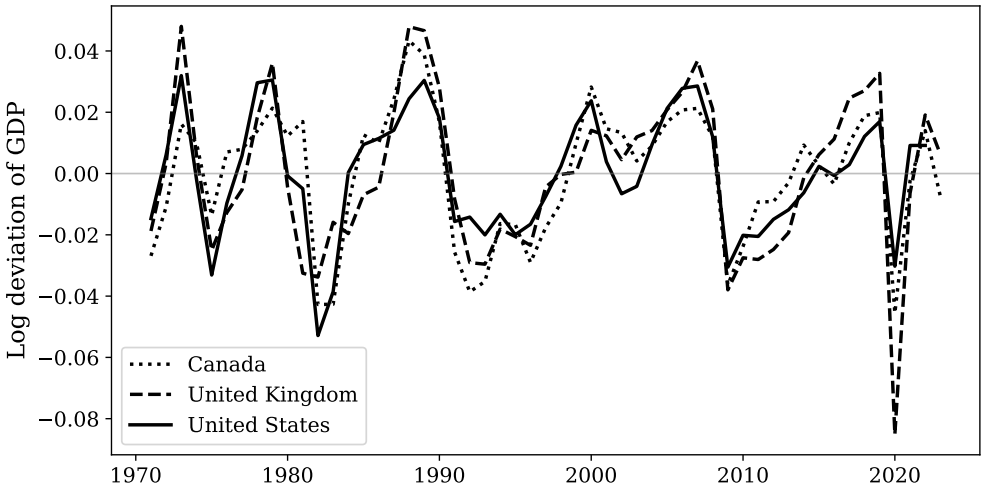


Figure 23.2: Log-deviations in GDP per capita (HP-filtered)

Notes: Hodrick-Prescott filtered log GDP, using $\lambda = 100$ for this annual data.

Table 23.2: Synchronized business cycles

Statistic	1971:Q3-2018:Q2	1971:Q3-2007:Q4
Correlations (US vs ROW)		
GDP	0.58	0.56
C	0.46	0.44
I	0.56	0.48
CPI	0.62	0.57
Backus-Smith Correlation (US vs ROW)		
Rel C vs RER	-0.49	-0.53

Notes: ROW includes Aus, Can, Fra, Ger, Ire, Ita, Jap, Swe, UK. All data detrended over 1971:Q3-2018:Q2 period using Hodrick-Prescott filter with $\lambda = 1600$. Source: update of Corsetti, Dedola, and Leduc (2008b).

Cross-border business-cycle correlations have long challenged the open-macro literature. At country level, TFP, one of the key drivers of business cycles, exhibits a low degree of synchronization.[10] In order to match business-cycle co-movements, quantitative models

[10]As shown by Backus, Kehoe, and Kydland (1992) and subsequent literature, joint estimation of TFP among advanced countries typically yields a positive but small covariance of contemporaneous shocks, and

therefore need to embed a sufficiently strong transmission mechanism for country-specific productivity innovations. We discuss the evidence on this mechanism at the end of this section, looking at the effects of identified TFP shocks in U.S. tradables production on macro variables, including output, in other G10 economies.[11]

International financial and economic integration puzzles. Before proceeding, we highlight four "puzzles" concerning the role of financial and goods-market integration in shaping the international business cycle.

A long-standing debate in open macro concerns the high correlation between national saving and investment. From the perspective of a world with perfect capital mobility, this appears counterfactual: investment should be higher in countries where returns are high—with capital flows naturally supplementing national saving. This is dubbed the "Feldstein-Horioka puzzle." While a high correlation may be produced by shocks that move saving and investment in tandem, the puzzle is commonly used to highlight the incidence of frictions in international capital mobility and/or goods trade—both ultimately impeding intertemporal trade and risk sharing.

A related debate revolves around the "home bias in equity portfolios puzzle." This is based on evidence that investors exhibit a strong preference for domestic assets, despite opportunities for international diversification. The question is whether home bias is a feature of optimal portfolio allocation, as opposed to being driven by trading and information frictions.

By the same token, one may note that consumption patterns across countries are less correlated than predicted by models that assume frictionless international risk-sharing. This discrepancy, known as the "consumption correlation puzzle," also suggests the need to investigate the nature and incidence of frictions in risk sharing, as opposed to fundamental determinants of portfolio diversification (e.g., different preferences or risk exposures).

Lastly, trade within countries significantly exceeds international trade, even among highly integrated economic regions. This phenomenon, known as the "home bias in trade puzzle," requires models of international trade to incorporate and calibrate parameters that replicate the home bias in demand. Obstfeld and Rogoff (2001) provide a thorough discussion of each of these puzzles, and more.

23.2.2 Exchange rates and relative prices

We now introduce international relative prices. While the analysis in this chapter is developed in real terms, it is useful to start with the definition of the nominal exchange rate (NER), denoted by $\mathcal{E}_t$. The NER represents the relative price of currencies, defined as the number of home currency units required for one unit of foreign currency. For example, when the home country is the U.S. and the foreign country is the EU, our notation means that one euro can be bought with $\mathcal{E}_t$ U.S. dollars. An *increase* in $\mathcal{E}_t$ represents a nominal *depreciation*

a positive but moderate spillover coefficient, by which positive innovations in a country drive TFP abroad with a delay.

[11] As further discussed below, the literature distinguishes between tradable goods, which can be exported and imported by a country, and non-tradable goods. The latter are generally defined as goods that are not traded internationally, due to prohibitively high transportation costs relative to their value added, (see, e.g., Obstfeld and Rogoff, 1996, p.199).

of the dollar (home currency), because more dollars are necessary to buy one euro, whereas a *decrease* in $\mathcal{E}_t$ indicates a dollar *appreciation*. In contrast with *bilateral* exchange rates, *trade-weighted* exchange rates (also known as *effective* exchange rate) offer a broader measure reflecting the value of a currency relative to a basket of foreign currencies (e.g., taking into account the yen, the renminbi, etc.).

The *Law of One Price* (LOOP) states that a good should cost the same abroad as at home when expressed in a common currency. Formally, LOOP holds for a specific good i if

$$P_{i,t} = \mathcal{E}_t P^*_{i,t},$$

where $P_{i,t}$ is the price of good i in dollars at time t and $P^*_{i,t}$ the price of the same good in euros. Hence, LOOP implies that the cost of the good in euros multiplied by the NER must equal the cost in dollars. For example, if a computer can sell at $P_{i,t}$ dollars in the U.S. and $P^*_{i,t}$ euros in Europe, a U.S. producer can sell it in Europe, exchange the euros for dollars and receive $\mathcal{E}_t P^*_{i,t}$ dollars. While LOOP generally holds well for tradable goods, such as commodities and luxury consumer items, it often breaks down for non-tradable goods, like consumer services, housing, transportation and utilities (Schmitt-Grohé, Uribe, and Woodford, 2022).

Purchasing Power Parity (PPP) is a similar concept, said to hold when the cost of a representative basket of goods in each country is identical. That is, PPP extends LOOP from individual goods to a broad basket of goods representative of households' actual consumption, reflecting the relative cost of living between countries. Mathematically, it is expressed as

$$P_t = \mathcal{E}_t P^*_t.$$

Note that if consumption baskets are the same in both countries, LOOP implies PPP. In general though, there are large deviations from absolute PPP in the data. Deviations from PPP arise from factors such as home bias (the difference in baskets), the existence of non-tradable goods and LOOP violations for traded goods. In the model that we present later, we focus on home bias as the main source of discrepancy between LOOP and PPP.

The *real exchange rate* (RER), denoted by $\mathcal{Q}_t$, reflects the relative cost of goods between a foreign and the home country, capturing how the price of a representative basket of goods abroad evolves over time relative to a basket of goods at home. It is defined as the nominal exchange rate adjusted for relative price levels at home P_t and abroad P^*_t,

$$\mathcal{Q}_t = \frac{\mathcal{E}_t P^*_t}{P_t}.$$

The representative basket at home costs P_t dollars and the representative basket abroad (which may have different components) P^*_t euros (or $\mathcal{E}_t P^*_t$ dollars). An increase in $\mathcal{Q}_t$ means that it becomes cheaper (in dollars) to buy the home basket compared to the foreign basket. In other words, the home currency can buy relatively less of the representative basket abroad. An increase in $\mathcal{Q}_t$ therefore reflects a *real depreciation* for the home country. As the equations above show, *absolute PPP* holds when $\mathcal{Q}_t = 1$.[12]

Most studies of PPP emphasize changes in the RER rather than its absolute level. This focus is primarily because changes in the RER can be calculated using consumer price indices

[12]Note that whereas we can use the nominal exchange rate $\mathcal{E}_t$ with a unit ("$\mathcal{E}_t$ dollars"), the real exchange rate does not have a unit (it is a relative cost of two baskets, both measured in a common currency).

(CPIs), which are widely available for many countries at relatively high frequencies, typically on a monthly basis. While the CPI provides information about the price of a basket of goods only up to a scalar, it effectively measures how the price of the basket changes over time. Mathematically, the RER can be decomposed into (log) changes in the NER and inflation differentials between countries:

$$\Delta q_t = \Delta e_t + \pi_t^* - \pi_t,$$

where $\Delta q_t = \ln \mathcal{Q}_t - \ln \mathcal{Q}_{t-1}$, $\Delta e_t = \ln \mathcal{E}_t - \ln \mathcal{E}_{t-1}$, and $\pi_t^{(*)} = \log P_t^{(*)} - \log P_{t-1}^{(*)}$ is home (foreign) CPI inflation. *Relative PPP* holds if $\Delta q_t = 0$ or $\Delta e_t = \pi_t - \pi_t^*$; that is if the depreciation rate of a country's currency against the domestic currency, Δe_t, matches the inflation differential between the foreign country and the domestic one.

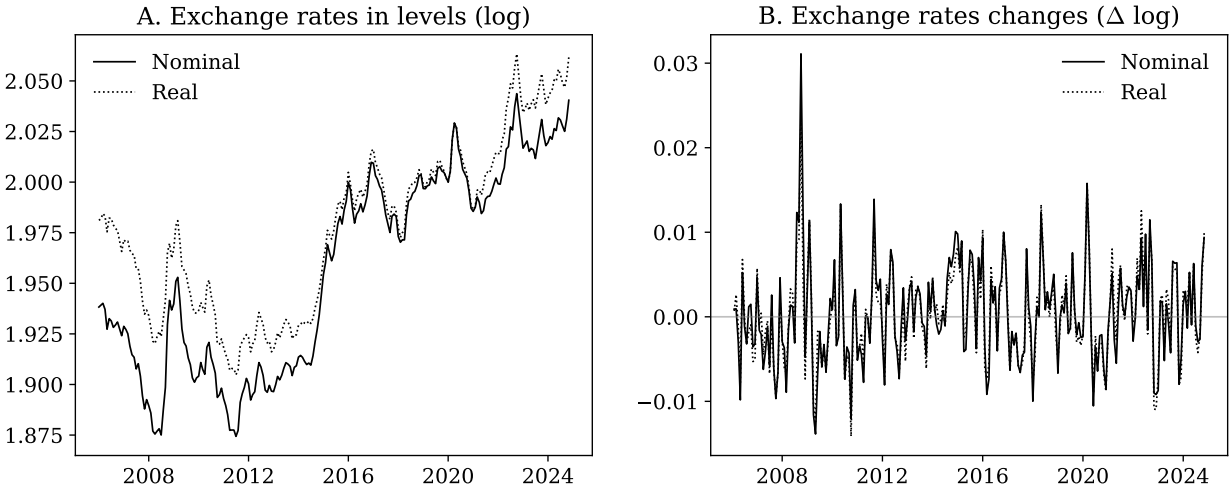


Figure 23.3: Nominal and Real U.S. Exchange Rates, in levels (left) and changes (right)

Source: FRED. The NER corresponds to the series DTWEXBGS in FRED, "Nominal Broad U.S. Dollar Index, Index Jan 2020=100, Monthly, Not Seasonally Adjusted." The RER corresponds to the series RBUSBIS, "Real Broad Effective Exchange Rate for United States, Index Jan 2020=100, Monthly, Not Seasonally Adjusted."

In the data, particularly for major currencies, the RER closely tracks the NER, as Figure 23.3 shows. In general, exchange rates in nominal and real terms are more volatile than macroeconomic aggregates, but less volatile than other financial market variables. Moreover, exchange-rate fluctuations exhibit no robust contemporaneous correlation with macroeconomic fundamentals (such as the money supply, interest rates and output levels). This fact, which has been labeled the "exchange-rate disconnect", has long challenged both macroeconomic and financial models. In the same vein, Meese and Rogoff (1983) famously established that empirical exchange-rate models cannot outperform the random walk in out-of-sample forecasting of exchange rates—the "Meese and Rogoff puzzle."

The exchange rate also displays a high degree of persistence, with weak mean reversion, as the left panel of Figure 23.3 demonstrates. This slow adjustment is commonly referred to as the "PPP puzzle." The right panel of the figure highlights the substantial volatility

of the RER, indicating that PPP does not hold in the short run. However, the tendency of RER changes to center around zero suggests that PPP holds in the long run.[13] Studies have shown that the RER is almost an order of magnitude more volatile than macroeconomic fundamentals such as inflation, consumption, and output. Moreover, it tends to be negatively correlated (if only weakly) with the ratio of domestic to foreign consumption. A negative correlation between the RER and the home-to-foreign relative consumption is puzzling: when the RER is appreciated, home goods are relatively expensive; if this reflects scarcity of the home good, one would expect home consumption to be lower, not higher, than foreign—the "Backus-Smith puzzle." For the U.S., the puzzle is illustrated by Figure 23.4, with a correlation between RER and relative consumption of around −0.5 (see also bottom row of Table 23.2).[14]

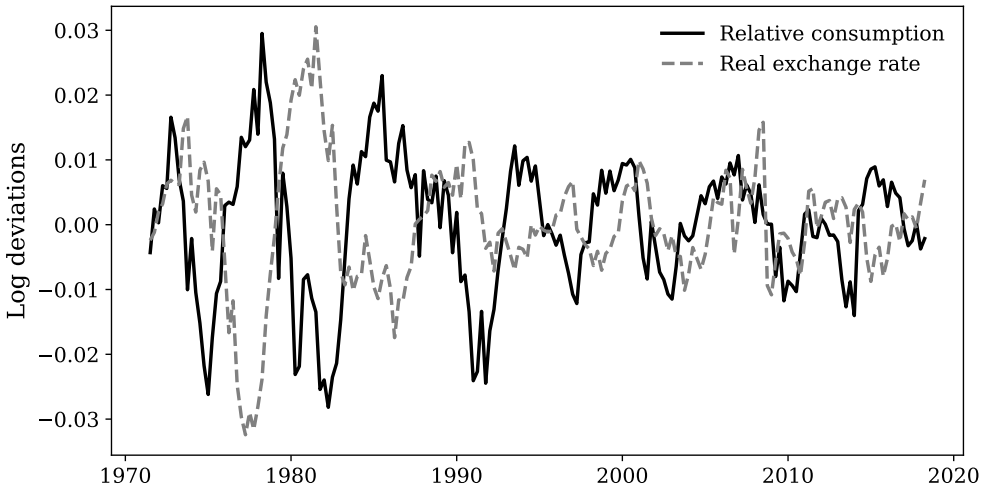


Figure 23.4: The U.S. real exchange rate and relative consumption (HP-filtered)

Source: update of Corsetti et al. (2008b). U.S. real exchange rate and relative consumption constructed relative to remaining PPP-weighted G10 economies. Hodrick-Prescott filtered series, using $\lambda = 1600$ for this quarterly data.

The "terms of trade" (TOT), denoted by $\mathcal{T}_t$, is the final important international relative price. It measures the relative price of the import basket ($P_{F,t}$, price of imported foreign goods at home) to the export basket ($P^*_{H,t}$, price of exported home goods abroad):

$$\mathcal{T}_t = \frac{P_{F,t}}{P^*_{H,t}\mathcal{E}_t},$$

[13] This is also true if the RER are computed using PPI- or wage-based price indexes. Empirical studies suggest that the RER is stationary, suggesting that despite large fluctuations at business-cycle frequencies, there is some version of PPP anchoring its long-run level (Taylor, 2002; Taylor and Taylor, 2004).

[14] The "Mussa Puzzle," first documented by Mussa (1986), highlights a dramatic increase in the volatility of nominal and real exchange rates following the transition of industrial countries from fixed to floating exchange rate regimes after the collapse of the Bretton Woods System in 1973. The rise in volatility cannot fully be explained by changes in domestic price levels, suggesting a disconnect between exchange rate movements and inflation. Since we focus on a real economy in this chapter, we will not have much to say about this.

where $P_{F,t}$ ($P^*_{H,t}$) is the home (foreign) import price index in local currency (that is, $P_{F,t}$ is measured in dollars and $P^*_{H,t}$ is measured in euros). The TOT and RER differ in that the import basket is different from the representative consumption basket at home and the export basket is different from the representative consumption basket abroad. In other words, while the RER reflects deviations in purchasing power, the TOT captures actual relative prices of traded goods. Empirically, the RER and TOT are only weakly correlated, with the TOT being less volatile.

23.2.3 Cross-country transmission of productivity shocks

We conclude this section with conditional evidence, which sets the stage for the rest of the chapter. Specifically, we complement the business-cycle statistics discussed so far with evidence on the cross-country transmission of technology improvements, highlighting how advancements in one country affect relative quantities, prices, and external balances. Figure 23.5 presents estimated impulse responses from Corsetti, Dedola, and Leduc (2014), which empirically examine the cross-country transmission of U.S. supply-side shocks relative to the rest of the world. This analysis employs a sign-restricted Bayesian vector auto-regression methodology. The figure illustrates the responses of U.S. relative quantities (manufacturing output and consumption), relative prices (real exchange rate and terms of trade), and external balances (net exports and net foreign assets) to a temporary increase in U.S. labor productivity in the tradable sector, *vis-à-vis* other G10 economies.

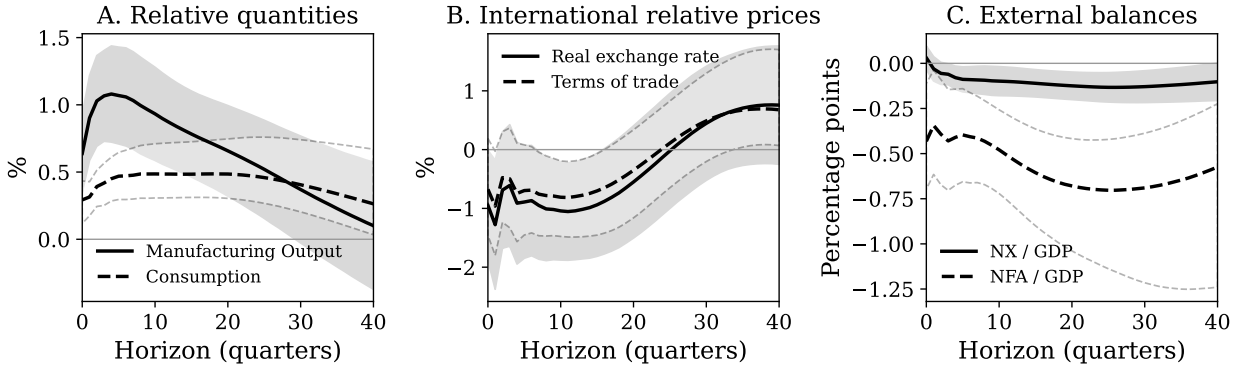


Figure 23.5: Empirical estimates of the cross-country transmission of productivity shocks: U.S. to rest of world

Source: Corsetti, Dedola, and Leduc (2014) (Figures 1 and 2). *Notes*: estimated impulse response to a positive U.S. labor productivity shock. Charts report median response of U.S. variables relative to G10 rest of the world, as well as the 16th and 84th percentiles of the posterior distribution, satisfying the sign restrictions detailed in Section 2.1 of Corsetti et al. (2014). Rest of the world is PPP-weighted and comprises the following advanced economies: Japan, Germany, UK, Italy, France, Canada, Australia, Sweden, and Ireland. RER = real exchange rate; TOT = terms of trade; NX/GDP = net exports to GDP; NFA/GDP = net foreign assets to GDP. Estimation sample: 1973Q1-2004Q4.

Although the U.S. and other G10 business cycles are to a large extent synchronized, the estimated responses reveal that a relative productivity improvement in the U.S. leads to an

increase in U.S. tradable output *relative* to the rest of the world; this increase in relative output goes hand in hand with an increase in relative consumption and appreciation of the domestic real exchange rate (the ratio of rest-of-the-world to U.S. consumer prices declines) and TOT, such that imported goods are relatively cheaper. External balances—net exports and net-foreign-asset positions—are counter-cyclical; conditional on the shock, the domestic economy imports and consumes more today, funded by international borrowing.

These patterns have been further explored by subsequent literature, relying on different methodologies and/or corroborating analyses of identified contemporaneous, *but persistent*, productivity shocks with analysis identifying 'news shocks' to productivity (see, e.g., Nam and Wang, 2015; Chahrour, Cormun, De Leo, Guerrón-Quintana, and Valchev, 2024). In the closed-economy literature, leading studies developed after the Covid-19 pandemic have recently elaborated on the notion of "Keynesian Supply shocks" (Guerrieri, Lorenzoni, Straub, and Werning, 2022), similar to the mechanism explained in this chapter (i.e., the idea that productivity shocks move both supply and demand, and under certain conditions, the demand effects can be dominant).

The pattern identified by the empirical estimates highlight the two key stylized facts that challenge open-macro theory: (1) net exports and the current account (external saving) tend to be counter-cyclical; and (2) international relative prices (i.e., the RER and TOT) tend to appreciate when output and consumption are relatively high. In what follows, we show how to bring modern international business-cycle theory closer in line with this evidence.

23.3 The workhorse open-economy model

We do so by presenting a two-country macro model, widely used to study the cross-border transmission of country-specific shocks. Our focus will be on productivity improvements. Throughout, we will confront the model with the stylized facts presented in the previous section—namely the pro-cyclicality of external deficits, and the co-movements of output, demand and exchange rates. In this section, we simplify our analysis by modeling endowment economies; in the next, we generalize our results to economies with endogenous production; in the third, we synthesize the main conclusions with three numerical exercises.

23.3.1 Model setup: preferences and technology

The model world economy consists of two countries: home H and foreign F, each populated by a continuum of identical households with unit mass. Throughout, we label foreign variables with an asterisk. We assume, for simplicity, that countries are of equal size. A representative home household receives an exogenous endowment of (tradable) home goods, $Y_{H,t}$, while the foreign agent has an exogenous endowment of the foreign good, $Y^*_{F,t}$. For consumers, the home and foreign goods are imperfect substitutes, so there is trade in goods internationally. In the global equilibrium, the separate markets for home and foreign goods, along with the markets for internationally traded assets, must simultaneously clear.

Household utility. In each country, the representative household derives utility from consuming a combination of domestic and imported goods. In the home country, aggregate

consumption is denoted by C_t. The expected lifetime utility of the representative home household U_t is:

$$U_t = \mathbb{E}_t \left[\sum_{j=0}^{\infty} \beta^j u(C_{t+j}) \right] \tag{23.1}$$

where $\beta \in (0,1)$ is the discount factor and $u(C) = (C^{1-\sigma} - 1)/(1-\sigma)$ is instantaneous utility from consumption (with $\sigma > 0$ representing the coefficient of relative risk aversion). Preferences of the representative foreign consumer are defined analogously, with consumption C_t^*, expected lifetime utility U_t^*, as well as $\beta^* = \beta$ and $\sigma^* = \sigma$.

The final consumption basket C_t is a CES aggregator of home, $c_{H,t}$, and foreign goods, $c_{F,t}$, known as the 'Armington aggregator' (Armington, 1969), of the form:

$$C_t = \left[a_H^{\frac{1}{\phi}} c_{H,t}^{\frac{\phi-1}{\phi}} + a_F^{\frac{1}{\phi}} c_{F,t}^{\frac{\phi-1}{\phi}} \right]^{\frac{\phi}{\phi-1}}. \tag{23.2}$$

This allows for imperfect substitutability between domestic and foreign goods—encompassing both the case of complementarity and substitutability—as well as 'home bias' in consumption, in line with the stylized fact that a greater proportion of domestic expenditure is devoted to domestic goods. With $a_H + a_F = 1$, the parameter a_H (a_F) represents the steady-state share of the home (foreign) goods in home expenditure such that when $a_H \in (0.5, 1]$ we have home bias in consumption. The parameter $\phi > 0$ denotes the constant elasticity of substitution between home and foreign goods, with higher (lower) values representing greater substitutability (complementarity) between home and foreign goods. As a result, this parameter is often referred to as the 'trade elasticity'.[15]

Given separable and time-additive preferences in equation (23.1), we can solve the household optimization problem by separating its intertemporal and intratemporal dimensions. As shown below, the representative household's intertemporal preferences will pin down aggregate consumption and savings, once the structure of international financial markets is defined. Taking aggregate consumption C_t as given, one can determine the demand for home and foreign goods, and derive a welfare-based price index of consumption, solving an intratemporal optimization in each period.

Let $p_{H,t}$ and $p_{F,t}$ represent the price of home and foreign goods, respectively, in the home economy. Consider a level of expenditures $\mathbf{E}_t = P_t C_t$ on home and foreign goods

$$\mathbf{E}_t = p_{H,t} c_{H,t} + p_{F,t} c_{F,t}.$$

The intratemporal decision of households involves minimizing their expenditure with respect to $c_{H,t}$ and $c_{F,t}$ subject to the Armington aggregator.

Solving this, home demand functions for home and foreign goods are, respectively:

$$c_{H,t} = a_H \left(\frac{p_{H,t}}{P_t} \right)^{-\phi} C_t \quad \text{and} \quad c_{F,t} = a_F \left(\frac{p_{F,t}}{P_t} \right)^{-\phi} C_t, \tag{23.3}$$

[15] For given aggregate consumption C_t, one can show that home and foreign goods are substitutes (complements) when $\phi > 1$ ($\phi < 1$) by calculating the second-order cross-partial derivatives of equation (23.2), $(\partial C_t^2)/(\partial c_{H,t} \partial c_{F,t})$.

where the home CPI is defined as:

$$P_t = \left(a_H p_{H,t}^{1-\phi} + a_F p_{F,t}^{1-\phi}\right)^{\frac{1}{1-\phi}}. \tag{23.4}$$

The foreign CPI P_t^* is similarly defined, with $p_{H,t}^*$ and $p_{F,t}^*$ denoting the prices of home and foreign goods in the foreign economy.[16]

Market clearing. Given that the world is a closed endowment economy, equilibrium involves market clearing for each good: $Y_{H,t} = c_{H,t} + c_{H,t}^*$ and $Y_{F,t} = c_{F,t} + c_{F,t}^*$ for all t.

Exchange rates and relative prices. Since we abstract from nominal rigidities, we normalize $\mathcal{E}_t$ to 1 and focus, instead, on relative prices: the TOT and RER. We also assume that the LOOP holds: $p_{i,t} = \mathcal{E}_t p_{i,t}^* = p_{i,t}^*$ for $i = \{H, F\}$. The TOT from the perspective of the home country is then:

$$\mathcal{T}_t = \frac{p_{F,t}}{p_{H,t}}. \tag{23.5}$$

Hence, $\mathcal{T}_t$ goes up when foreign imports become relatively more expensive, corresponding to a worsening in the home TOT. From the perspective of consumers, taking the TOT as given, the intratemporal consumption demand for domestic and foreign goods $\{c_{H,t}, c_{F,t}, c_{H,t}^*, c_{F,t}^*\}$ arising from the ratio of the equations in (23.3) must satisfy:

$$\frac{1}{\mathcal{T}_t} = \left(\frac{a_H}{a_F}\right)^{\frac{1}{\phi}} \left(\frac{c_{H,t}}{c_{F,t}}\right)^{-\frac{1}{\phi}} = \left(\frac{a_H^*}{a_F^*}\right)^{\frac{1}{\phi}} \left(\frac{c_{H,t}^*}{c_{F,t}^*}\right)^{-\frac{1}{\phi}}. \tag{23.6}$$

Given the normalization of $\mathcal{E}_t = 1$, the RER $\mathcal{Q}_t$ satisfies

$$\mathcal{Q}_t = \frac{\mathcal{E}_t P_t^*}{P_t} = \frac{P_t^*}{P_t}. \tag{23.7}$$

An increase in $\mathcal{Q}_t$ is a home real depreciation, with foreign consumption becoming more expensive relative to home consumption. Importantly, although the LOOP is assumed to hold for each tradable good, the LOOP is not sufficient for the price of consumption bundles to be equal across countries (i.e., for PPP to hold). PPP requires that $P_t = \mathcal{E}_t P_t^* = P_t^*$, which would imply $\mathcal{Q}_t = 1$, a condition regularly rejected in the data, as explained in the previous section. An attractive feature of Armington aggregation is that absolute PPP will not generally hold. It will only hold if two conditions are satisfied: (i) LOOP holding in both goods markets, and (ii) consumption baskets being identical across countries.[17]

[16] To be clear, we use H to denote a specific good. This means it is "home-produced" for the domestic economy but actually "foreign-produced" from the perspective of the other country. Thus, a_H^* is the CES weight on the H good for the foreign economy. If countries are symmetric and have a stronger taste for their own good, then $a_H = a_F^* > a_H^* = a_F$.

[17] In addition to accounting for home bias in consumption, as we do here, one can also model deviations from PPP by: (i) modeling deviations from the LOOP, for example by allowing 'pricing-to-market' by firms, or (ii) allowing for non-tradable goods within the model, for which the LOOP will not necessarily hold.

By using the RER definition (23.7), home CPI (23.4) and the analogous expression for foreign CPI, we can express a relationship between the RER and TOT as:

$$\mathcal{Q}_t^{1-\phi} = \frac{a_H^* + a_F^* \mathcal{T}_t^{1-\phi}}{a_H + a_F \mathcal{T}_t^{1-\phi}}. \tag{23.8}$$

According to the model, the co-movement between the RER and TOT will depend on the degree of home bias. If $a_H = a_F^* > 0.5$, then the co-movement will be positive; while if $a_H = a_F^* = 0.5$ then PPP will hold and fluctuations in TOT will not translate into fluctuations in RER. Counterfactually, away from PPP, the model predicts a perfect correlation between $\mathcal{Q}$ and $\mathcal{T}$. Model specifications that allow for (plausible) differences between consumption baskets—e.g., via the presence of non-traded goods and distribution costs (see Benigno and Thoenissen, 2008; Corsetti, Dedola, and Leduc, 2008a)—can resolve this.

So far, we have not characterized the consumption-saving decision by households. Specifically, in open economies, since domestic aggregate consumption can deviate from domestic output via trade in financial assets (i.e., in the model, $Y_{H,t}$ need not equal C_t), the structure of financial markets is crucial for determining the global equilibrium, including the exchange-rate determination and capital flows across borders. To close the model, a specification for international financial markets is needed.

23.3.2 International financial markets and intertemporal choices

To understand the role of international financial markets, and their degree of 'openness', in shaping the global equilibrium, we contrast three benchmark cases (akin to those in Chapter 5), that differ in the extent to which country specific risks are insured (or 'traded'):

1. **Complete markets**: where all risk is traded via a complete set of Arrow-Debreu securities, and the RER is pinned down by the ratio of marginal utilities across borders.

2. **Incomplete markets**: where insurance is partial, so that equilibrium reflects non-traded risks:

   (i) **Financial autarky**: with no markets for international assets, the RER must adjust to balance international trade in goods.

   (ii) **Riskless bonds only**: with trade in a single non-contingent bond, through which households can self insure, the RER is pinned down by the 'uncovered interest parity' (UIP) condition.

We describe each in turn in the remainder of this sub-section. Throughout the analysis we assume that agents commit to repay bonds (regardless of whether they are contingent or not). The commitment assumption will be relaxed in Chapter 24.

Complete markets

Let s_t denote the time-t state of the world and suppose $s_t \in S$, where S is the exogenous part of the state space (i.e., values of the two endowments). If markets are complete—for instance,

if there is a full set of Arrow-Debreu securities for each potential state of the world—all risk is traded. The representative home household's intertemporal budget constraint is then:

$$P_t C_t + \int_{s_{t+1}} q_t(s_{t+1}) \mathbb{B}_{H,t}(s_{t+1}) \, \mathrm{d}s_{t+1} \leq \underbrace{\mathbb{B}_{H,t-1} + p_{H,t} Y_{H,t}}_{\text{current income}},$$

where $\mathbb{B}_{H,t}(s_{t+1})$ denotes the quantity of Arrow-Debreu securities paying one unit of the home consumption basket upon realization of the state s_{t+1} at time $t+1$, traded at the price $q_t(s_{t+1})$ at time t. For convenience, this notation suppresses the dependence on history before t (the whole history of shocks). For example, $q_t(s_{t+1})$ is the price at time t, given a history up until and including t, of a one-unit payoff of H goods at $t+1$, and for different histories this $q_t(s_{t+1})$ will in general take on different values.[18]

Maximizing the representative home household's expected lifetime utility subject to this budget constraint yields the following Euler equation for each state $s_{t+1} \in S$:

$$u'(C_t) \frac{q_t(s_{t+1})}{P_t} = \Pr(s_{t+1}|s_t) \beta u'(C_{t+1}(s_{t+1})) \frac{1}{P_{t+1}(s_{t+1})}, \tag{23.9}$$

where $\Pr(s_{t+1}|s_t)$ denotes the probability of transitioning from state s_t at time t to s_{t+1} at $t+1$.

Consider guaranteeing one unit of consumption good for a home investor at time $t+1$. For this, at period t, the investor has to buy $P_{t+1}(s_{t+1})$ units of the Arrow-Debreu security for each $s_{t+1} \in S$. This purchase would cost $q_t(s_{t+1}) P_{t+1}(s_{t+1})$ in terms of the home basket and in total $\int q_t(s_{t+1}) P_{t+1}(s_{t+1}) \mathrm{d}s_{t+1}$ units as a bundle. In terms of time-t consumption goods, this bundle costs $\int q_t(s_{t+1}) P_{t+1}(s_{t+1}) \mathrm{d}s_{t+1}/P_t$ units of the home consumption bundle today. Let us call the gross real return $R_t \geq 1$ the inverse of the cost of this bundle. That is, because $1/R_t$ units of the home consumption bundle today can guarantee one unit of the home consumption bundle tomorrow, it follows that one unit today would guarantee R_t units tomorrow. Using (23.9), that is

$$\frac{1}{R_t} = \frac{1}{P_t} \int q_t(s_{t+1}) P_{t+1}(s_{t+1}) \mathrm{d}s_{t+1} = \int \Pr(s_{t+1}|s_t) \beta \frac{u'(C_{t+1}(s_{t+1}))}{u'(C_t)} \mathrm{d}s_{t+1}.$$

In the foreign economy, there is equivalently a full set of Arrow-Debreu securities with prices q^*, and hence we can define R_t^* to be the number of units of the foreign consumption bundle obtained by a riskless investment of one unit of such a bundle today.[19] Thus the gross real returns R_t and R_t^* for the representative home and foreign households satisfy

$$1 = \beta R_t \mathbb{E}_t \left[\frac{u'(C_{t+1})}{u'(C_t)} \right] = \beta R_t^* \mathbb{E}_t \left[\frac{u'(C_{t+1}^*)}{u'(C_t^*)} \right]. \tag{23.10}$$

Now, consider the following: rather than investing in home Arrow-Debreu securities, first exchange one unit of the home consumption bundle into $1/\mathcal{Q}_t$ units of the foreign bundle.

[18] The notation based on Chapter 7 would therefore have been $q_t(S^{t+1})$, with $S^{t+1} = \{S^t, s_{t+1}\}$.

[19] The existence of Arrow-Debreu markets in both countries means an excess of assets—twice as many as we need to complete markets. We use this assumption for symmetry and will of course invoke the absence of arbitrage, as in Chapter 14, when needed.

Then invest it in the foreign Arrow-Debreu securities and receive $R_t^*/\mathcal{Q}_t$ units of foreign bundle at period $t+1$, which can be converted into $R_t^*\mathcal{Q}_{t+1}/\mathcal{Q}_t$ units of the home bundle. No-arbitrage implies

$$1 = \beta R_t \mathbb{E}_t \left[\frac{u'(C_{t+1})}{u'(C_t)} \right] = \beta R_t^* \mathbb{E}_t \left[\frac{u'(C_{t+1})}{u'(C_t)} \frac{\mathcal{Q}_{t+1}}{\mathcal{Q}_t} \right]. \tag{23.11}$$

The above is a core condition in international finance: the (real) Uncovered Interest Parity (UIP) condition. The expected rate of RER depreciation is a function of (real) interest differentials and expected growth of marginal utilities.

When markets are complete, no-arbitrage implies a stronger condition, holding state by state. Combining home and foreign state-by-state Euler equations yields

$$\frac{u'(C_t)}{u'(C_t^*)} \frac{P_t^*}{P_t} = \frac{u'(C_{t+1}(s_{t+1}))}{u'(C_{t+1}^*(s_{t+1}))} \frac{P_{t+1}^*(s_{t+1})}{P_{t+1}(s_{t+1})} \quad \forall\, t. \tag{23.12}$$

The above equation leads to the key implication of *full risk sharing* under complete markets. Define $t=0$ as the initial period when the economy starts with the complete set of financial contracts already in place. At any point in time after that, it must be the case that:

$$\frac{u'(C_0)}{u'(C_0^*)} \mathcal{Q}_0 \equiv \kappa_0 = \frac{u'(C_t(s_t))}{u'(C_t^*(s_t))} \mathcal{Q}_t(s_t) \quad \forall\, t.$$

With perfect insurance, the marginal-utility ratio across two countries is pinned down by the wealth distribution—reflected by the difference in consumption—at time 0, and remains constant forever. The initial distribution at time 0 is of course endogenous, as it depends on the equilibrium asset and goods prices, which are in turn a function of relative endowments.[20] If countries are assumed to be initially perfectly symmetric (as they typically are in the workhorse model we consider), then $\kappa_0 = 1$; otherwise it endogenously takes values above or below one that cannot be ignored in the computation of the equilibrium. However, the key is that under complete markets κ (whatever value it takes) is independent of time.

Using our specification of preferences, the *full risk-sharing* condition simplifies to:

$$\mathcal{Q}_t = \kappa_0 \left(\frac{C_t}{C_t^*} \right)^{\sigma} \quad \forall t. \tag{23.13}$$

Under complete markets, there is a tight equilibrium relation between relative marginal utilities of consumption and the RER, the implications of which we study in the next section.[21] The full risk-sharing condition states that home consumption can increase relative to foreign consumption only if the home country experiences a real depreciation (i.e., domestic consumption becomes relatively cheaper).[22] This outcome is ensured by *ex post* contingent

[20] See Corsetti, Lipinska, and Lombardo (2025) for an analysis of risk sharing that distinguishes between the effects on consumption smoothing and relative wealth, driven by the repricing of assets.

[21] The correlation between relative consumptions and RER is 1 according to the model, but is negative in the data (see Figure 23.4 and Table 23.2), illustrating the Backus-Smith puzzle.

[22] In this chapter we abstract from (time) preference shocks, under which the condition includes an additional term as a function of this shock.

transfers across borders. From a welfare perspective, these transfers are efficient: a low price indicates that the home consumption good is relatively more abundant. As long as there are no other distortions (e.g., nominal rigidities), the complete-market specification of our real economy is first-best from the perspective of a global planner.

Closed-form solutions for prices and quantities can be computed as follows. Within a given period, normalize prices to $P_t = 1$, take C_t and C_t^* as given, and then combine the demand functions (23.3) with market clearing conditions for each good. Together with (23.13), this leaves the same number of equations as unknowns, with κ_0 determined by relative wealth/income.[23]

Financial autarky

Define the trade balance as the value of exports minus value of imports. In units of home consumption this is:

$$tb_t \equiv \frac{p_{F,t}}{P_t} c_{F,t} - \frac{p_{H,t}}{P_t} c^*_{H,t}. \tag{23.14}$$

Under perfect risk sharing, any non-zero (real) trade balance would be entirely financed *ex post* by state-contingent payments from Arrow-Debreu securities. At the other extreme, with no trade in international financial markets, *financial autarky*, external trade must balance in each and every period (i.e., $tb_t = 0$):

$$\mathcal{T}_t c_{F,t} - c^*_{H,t} = 0 \quad \Longleftrightarrow \quad C_t = \frac{p_{H,t}}{P_t} Y_{H,t}. \tag{23.15}$$

The TOT adjusts in equilibrium to ensure balanced trade. As a result, the expected real depreciation no longer obeys the UIP condition (23.11). Since no risk can be insured through financial contracts, in general any shock will drive a wedge in marginal utilities between home and foreign countries. Without the ability to transfer wealth intertemporally, in effect, the model is the two-country counterpart to closed-economy models with hand-to-mouth consumers. However, in open economies, consumption and output comprise different goods, so that real income crucially depends on the TOT, since from (23.4) we have

$$\left(\frac{p_{H,t}}{P_t}\right)^{1-\phi} = \frac{1}{a_H + (1 - a_H)\mathcal{T}_t^{1-\phi}}.$$

Under financial autarky, home households face a static maximization each period:

$$\max_{C_t} \; u(C_t) \quad \text{subject to} \quad P_t C_t = p_{H,t} Y_{H,t}.$$

Therefore, this case can be solved period by period, since there are no borrowing and saving mechanisms.[24]

When markets are incomplete, it turns out to be useful to define a notion of relative wealth levels of the two countries, $\mathcal{W}_t$, as many of the key equilibrium objects depend on how this

[23] In the complete-markets model, the relevant initial wealth levels are total present-value wealth amounts; under financial autarky, discussed next, they are given by $p_{H,t}Y_{H,t}$ and $p_{F,t}Y^*_{F,t}$, respectively.

[24] Note, of course, that domestic and foreign interest rates can be computed as the values that clear the respective saving markets at every period.

notion evolves over time. $\mathcal{W}_t$ is constant over time in the complete-markets model but will evolve stochastically over time in the autarky model and with riskless bonds only. To begin, let us denote by $\chi_t \equiv u'(C_t)/P_t$ the Lagrange multiplier on the budget constraint above. This variable can be interpreted as the marginal utility of increased consumer resources (measured in units of the home consumption basket) and, as the box below argues, will be strictly decreasing in the consumer's current wealth. Given that, we define

$$\mathcal{W}_t \equiv \frac{\chi_t^*}{\chi_t} = \frac{u'(C_t^*)/P_t^*}{u'(C_t)/P_t} = \left(\frac{C_t}{C_t^*}\right)^{\sigma} \frac{1}{\mathcal{Q}_t}. \tag{23.16}$$

Recall our definition of the constant κ under complete markets. We see from the above that $\mathcal{W}_t$ is its inverse (i.e., $1/\kappa_t = \mathcal{W}$), but away from complete markets nothing insures that the ratio of Lagrange multipliers remains constant over time, i.e., non-traded risk causes the analog of κ_0 under complete markets to become state contingent and time varying ($\kappa_t(s)$). As such, it can be interpreted as capturing, at each point in time, the *wealth gap* created by uninsured risk, relative to complete markets.

Lagrange multiplier as the marginal utility of income (wealth)

Consider a static model where a consumer is choosing a vector of consumption goods to maximize $u(C)$, where C is a CES aggregate of goods, subject to a budget $PC = Y$. Here, Y is total income, in units of the CES aggregate, and P is the price defined as the smallest cost of obtaining one unit of the CES aggregate. That is, we have a static version of the model described in equations (23.1)–(23.4) above. For this problem, we can write down the Lagrangian

$$\mathcal{L} \equiv u(C) + \chi\left(Y - PC\right).$$

The first-order condition is $u'(C) = \chi P$, as in the text. Moreover, quite trivially, the maximum utility equals $u(Y/P)$, which means that the derivative of maximized utility with respect to income, Y, equals $u'(Y/P)/P = \chi$. Given a strictly concave utility function u, therefore, χ is strictly decreasing in income.

The interpretation that the Lagrangian multiplier represents a marginal utility of wealth is also a general result for any utility maximization problem with a standard budget constraint. In a dynamic model, as in the main text, Y/P needs to be replaced by period-t resources minus saving. Thus, given prices, χ_t would represent (the inverse of) this wealth measure. Under financial autarky, that wealth measure is simply the current world market value of the home endowment; in the bond economy, it is the amount the consumer chooses to optimally consume, which in turn is increasing in consumer wealth due to consumption being a normal good.

Hence $\mathcal{W}_t$ represents how rich the home country is compared to the foreign country in equilibrium. When $\mathcal{W}_t > 1$, the home country is wealthier than the foreign country at period t, which is also reflected in a change in the relative demands for the two goods.

Riskless bonds only

In our final case, we model a particular form of market incompleteness, allowing for cross-border trade in discount bonds—a common approach for capturing financial-market incompleteness in the literature. To keep the analysis simple, and relying on our normalization $\mathcal{E}_t = 1$, we posit trade in a single nominal one-period bond B_t, with nominal gross return $(1 + i_t)$ and price $Q_t = 1/(1 + i_t)$. The budget constraint of the home household is now

$$C_t = \frac{p_{H,t}Y_{H,t} - (Q_t B_t - B_{t-1})}{P_t}.$$

Current consumption can now be financed not only out of the domestic endowment, but also by borrowing (for $B_t < 0$). Therefore, in equilibrium, the trade balance will be $(Q_t B_t - B_{t-1})/P_t$. The current account, defined as the change in the net foreign asset position of the country (in our model, $ca_t = (B_t - B_{t-1})/P_t$), results from the sum of the trade balance and the earnings on net foreign assets. A positive current-account balance means that the home country is accumulating net claims *vis-à-vis* the foreign one.

The household maximizes expected discounted lifetime utility (23.1) subject to

$$p_{H,t+j}Y_{H,t+j} + B_{t+j-1} - P_{t+j}C_{t+j} - Q_{t+j}B_{t+j},$$

a constraint we associate with the Lagrange multiplier χ_{t+j}, as above.[25] This problem yields the following Euler equation for the home household:

$$Q_t = \beta \mathbb{E}_t \left[\frac{u'(C_{t+1})}{u'(C_t)} \frac{P_t}{P_{t+1}} \right] \tag{23.17}$$

and analogously for the representative foreign household (where again we have that $\chi_t = u'(C_t)/P_t$). Under our assumption that there is trade in one asset only, we can always define the home real interest rate as the expected marginal utility growth

$$R_t = \left(\beta \mathbb{E}_t \left[\frac{u'(C_{t+1})}{u'(C_t)} \right] \right)^{-1},$$

Because of our assumption that the NER is identically equal to one, for no arbitrage in the bond market, $Q_t = Q_t^*$ has to hold. Thus,

$$\mathbb{E}_t \left[\frac{u'(C_{t+1})}{u'(C_t)} \frac{P_t}{P_{t+1}} \right] = \mathbb{E}_t \left[\frac{u'(C^*_{t+1})}{u'(C^*_t)} \frac{P^*_t}{P^*_{t+1}} \right], \tag{23.18}$$

which is a version of the real UIP condition. Using the definition of real interest rates, it can be rewritten as follows:

$$(R_t)^{-1} \mathbb{E}_t \left[\frac{P_t}{P_{t+1}} \right] + \mathrm{cov}_t \left(\frac{u'(C_{t+1})}{u'(C_t)}, \frac{P_t}{P_{t+1}} \right) = (R^*_t)^{-1} \mathbb{E}_t \left[\frac{P^*_t}{P^*_{t+1}} \right] + \mathrm{cov}_t \left(\frac{u'(C^*_{t+1})}{u'(C^*_t)}, \frac{P^*_t}{P^*_{t+1}} \right). \tag{23.19}$$

[25] Implicit here is also a borrowing constraint that will not bind in the approximate solution studied below.

This relates real interest rate differentials to expected inflation differentials. Under our normalization of the NER (equal to one), these are tightly linked to expected real depreciation.

A key difference between this 'bond economy', relative to complete markets, is that the state-by-state condition (23.12) no longer holds. With trade in a single bond, shocks still create an *ex-post* marginal-utility wedge, reflecting risk that cannot be traded by intertemporal smoothing, namely:

$$\epsilon_{t+1} \equiv \frac{u'(C_{t+1})}{u'(C_t)}\frac{P_t}{P_{t+1}} - \frac{u'(C^*_{t+1})}{u'(C^*_t)}\frac{P^*_t}{P^*_{t+1}} = \frac{\chi_{t+1}}{\chi_t} - \frac{\chi^*_{t+1}}{\chi^*_t}. \tag{23.20}$$

From the real UIP condition, we see that this wedge must equal zero in expectation: the relative evolution of wealth is random with zero drift in expectation.

23.3.3 Solving the model

Allowing shocks to emanate from country endowments, we define equilibrium as prices and quantities satisfying the key conditions above: consumer maximization for the two countries and market clearing.

Log-linearization. We log-linearize the model around its symmetric non-stochastic steady state with $\overline{\mathcal{Q}} = \overline{TOT} = 1$, imposing home bias in preferences ($a_H = a^*_F > 1/2$). From hereon, steady-state variables will be denoted with a bar (i.e., $\overline{x}$) and percent deviations from the steady state will be labelled with a hat (i.e., $\hat{x}_t = (x_t - \overline{x})/\overline{x}$).

Two key log-linearized expressions will be used recurrently in the derivation of the equilibrium below. The first links the RER to the TOT (23.8):

$$\widehat{\mathcal{Q}}_t = (2a_H - 1)\widehat{\mathcal{T}}_t. \tag{23.21}$$

Absent home bias ($a_H = 1/2$), PPP holds and $\widehat{\mathcal{Q}}_t = 0$ independently of equilibrium fluctuations in the TOT.

The second expression characterizes how the wealth gap $\mathcal{W}_t$ (23.16) can be written as a function of the consumption gap and the real exchange rate:

$$\hat{\mathcal{W}}_t = \sigma(\hat{C}_t - \hat{C}^*_t) - \widehat{\mathcal{Q}}_t, \tag{23.22}$$

an expression we have seen must be zero under full risk sharing. Using the above and taking expectations of the first-order approximation of (23.20), we see that

$$\mathbb{E}_t[\hat{\mathcal{W}}_{t+1} - \hat{\mathcal{W}}_t] = \mathbb{E}_t[\hat{\epsilon}_{t+1}] = 0.$$

In a bond economy, shocks result in the accumulation (or reduction) of net foreign assets. Hence, relative wealth will change in response to shocks at each point in time. This condition also implies that, in the linearized version of the bond economy, relative wealth is non-stationary: the economy does not converge back to the symmetric steady state around which the equilibrium is approximated. The non-stationarity of the wealth distribution is due to the fact that the model is linearized around a non-stochastic steady state in which the real

interest rate is equal to the inverse of the discount factor ($\beta R = 1$). The true, non-linear solution of the model does not, however, exhibit non-stationarity: in response to shocks, the economy is drifting slowly around an average relative wealth position of zero but never settles at zero since shocks are reoccurring and are large enough to influence all prices.[26] In Chapter 11, we studied closed-economy heterogeneous-agent economies and one can imagine extending the present model to include a continuum of countries (and a continuum of goods), rather than just two (of each). A law of large numbers for the productivity shocks, which would then be idiosyncratic and hence 'small', would then imply the existence of a steady state where $\beta R < 1$.

Since the log-linearized setting is a very convenient tool, we will instead slightly change the model so as to yield convergence back to steady state. This is accomplished with the aid of 'Uzawa-style' preferences, where the discount factor is assumed to rises with the accumulation of net debt.[27]

23.3.4 Global equilibrium: a relative demand-relative supply framework

In the rest of this section, we study the properties of the model, discussing how it can be brought in line with the empirical responses shown in Figure 23.5. At an aggregate level, global output—which we treat as exogenously given for now—must be equal to global consumption, independently of the structure of financial markets:

$$\hat{Y}_{H,t} + \hat{Y}^*_{F,t} = \hat{C}_t + \hat{C}^*_t.$$

The equilibrium allocation of consumption across borders and international relative prices will nonetheless vary with the degree of risk sharing. For this reason, it is instructive to solve and study the model in relative terms, showing first how the demand for the goods produced in each economy vary as a function of the TOT (the relative price of output), and the wealth gap. Using this schedule in conjunction with relative supply, we then characterize the equilibrium allocation as a function of fundamental (supply) shocks, and analyze their international propagation.

Relative demand. We now introduce a notion of *relative demand* for home vs. foreign goods as a function of wealth gap and international relative prices. Define the total demands for each good by $D_{H,t} \equiv c_{H,t} + c^*_{H,t}$ and $D^*_{F,t} \equiv c_{F,t} + c^*_{F,t}$. Combining this definition with (23.6), showing how the ratio of the home to foreign demand for Home (Foreign) goods depend on the TOT, and (23.8), showing how the TOT are related to the RER, we obtain the following expression for the (log-linearized) relative total demand

$$\hat{D}_{H,t} - \hat{D}^*_{F,t} = (2a_H - 1)\left(\hat{C}_t - \hat{C}^*_t\right) + 4a_H(1 - a_H)\phi\widehat{\mathcal{T}}_t. \tag{23.23}$$

[26] The borrowing constraints will prevent relative wealth from exploding.

[27] Alternatively, one can assume infinitesimal costs of holding foreign assets, sufficient to motivate agents to run their external assets down in the long run. Key differences across solutions are analyzed by Schmitt-Grohé and Uribe (2003) and Bodenstein (2011).

Differences in demand for home goods and foreign goods come from two sources: (i) differences in the level of consumption, driving the demand for each goods depending on home bias, and (ii) international relative prices. We know, from (23.22), that the consumption difference comes from (i) difference in wealth and (ii) relative price level of each country. Using (23.22), we can obtain:

$$\hat{D}_{H,t} - \hat{D}^*_{F,t} = \sigma^{-1}(2a_H - 1)\hat{\mathcal{W}}_t + \sigma^{-1}\left[4a_H(1-a_H)\left(\sigma\phi - 1\right) + 1\right]\widehat{\mathcal{T}}_t, \tag{23.24}$$

This expression highlights the two key channels affecting relative demand for home and foreign goods. For given international relative prices $\widehat{\mathcal{T}}$, the wealth gap drives relative demand in proportion to the degree of home bias a_H—as captured by the coefficient $(2a_H - 1)$. For a given wealth gap $\hat{\mathcal{W}}_t$, a fall in the relative price of home goods (higher $\widehat{\mathcal{T}}$) unambiguously increases the relative demand for home goods.[28] This *substitution effect* of relative price movements is always negative, since relative demand of home goods falls with an increase in its relative price (the inverse of the terms of trade).

This highlights a crucial result: substitution is the only active channel either when markets are complete ($\hat{\mathcal{W}}_t = 0$), in which case the adjustment of prices supports the first-best allocation, or if there is no home bias ($a_H = 1/2$) irrespective of the degree of risk sharing. With no home bias, the RER is constant, so relative demand is independent of transfers of purchasing power across borders.[29] With home bias, under incomplete markets, wealth effects associated with non-traded risk can work against substitution effects (in contrast with the first best and thus inefficiently) as a worsening in the TOT (i.e., a drop in the price of a country goods supply) pushes down on (imperfectly insured) households' purchasing power.

Relative supply. The relative supply of the home and foreign goods is independent of relative prices—simply $\hat{Y}_{H,t} - \hat{Y}^*_{F,t}$ here as we, for now, work with an endowment economy.

Equilibrium. The equilibrium is obtained combining the two schedules. To gain insights on its properties, the case of financial autarky is particularly instructive, since here the relative wealth gap is tightly linked to current real incomes and thus relative prices. An expression for the wealth gap under financial autarky can be derived using equation (23.22) and log-linearizing (23.15), which invokes equilibrium in that it uses the autarky budget constraint, along with its foreign analog. Evaluated at a given supply of home and foreign goods, the wealth gap can be written as follows:

$$\hat{\mathcal{W}}^{FA}_t = \sigma\left[\hat{Y}^S_{H,t} - \hat{Y}^{*S}_{F,t} - 2(1-a_H)\widehat{\mathcal{T}}_t\right] - (2a_H - 1)\widehat{\mathcal{T}}_t. \tag{23.25}$$

Substituting this into (23.24), and equating demand and supply, the relative demand schedule can be written as:

$$\hat{D}_{H,t} - \hat{D}^*_{F,t} = \left[1 - 2a_H\left(1 - \phi\right)\right]\widehat{\mathcal{T}}_t. \tag{23.26}$$

Under financial autarky, the coefficient on $\widehat{\mathcal{T}}_t$ will unambiguously be positive if PPP holds (i.e., if $a_H = \frac{1}{2}$), as is also the case under complete markets, such that relative home demand

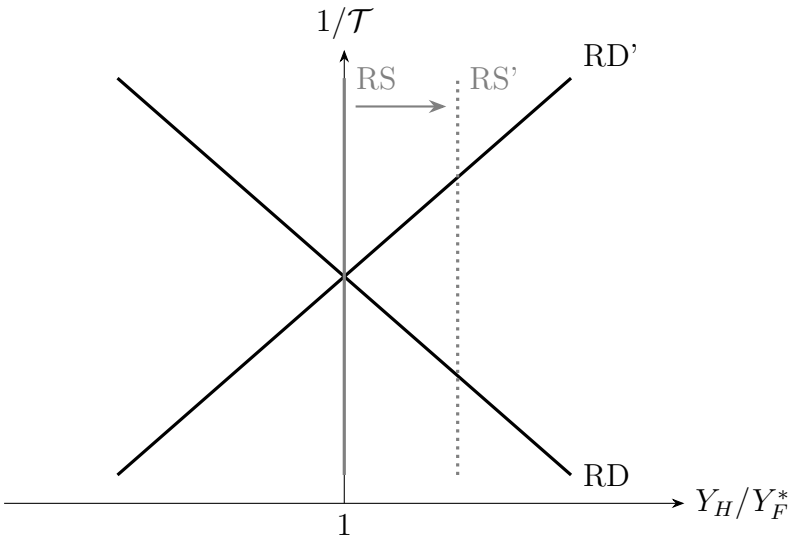
[28] Note that the coefficient on $\widehat{\mathcal{T}}_t$ in brackets is unambiguously positive, and increasing in the trade elasticity ϕ.

[29] See the controversy between Keynes and Ohlin on the effects of 'transfers' in the 1920s.

is unambiguously downward sloping in the international relative price of home output. But, under financial autarky, relative demand can be negatively or positively associated with relative price when there is home bias, depending on the elasticity ϕ. Here, relative demand is ultimately shaped by the endogenous response of the wealth gap to shocks. For example, the equilibrium relative demand for the home goods will be increasing in its relative price (i.e., $\frac{\partial(\hat{D}_{H,t}-\hat{D}^*_{F,t})}{\partial \mathcal{T}_t} < 0$) with an elasticity $\phi < 1 - \frac{1}{2a_H}$. Intuitively, with home bias in consumption and strong complementarity between home and foreign goods, home demand needs to rise enough to buy the higher supply of home goods. This is only possible if home goods prices rise in equilibrium, driving up home incomes.

Figure 23.6 illustrates the equilibrium by plotting relative demand (RD) and relative supply (RS) in the space $\{Y_H/Y_F, 1/\mathcal{T}\}$—note that y-axis reports the inverse of the TOT. RD denotes the relative demand schedule for $\phi > 1-\frac{1}{2a_H}$, upward sloping in this space, while RD' represents relative demand with sufficient complementarity which slopes downward. Relative supply in the endowment economy is a vertical line. The figure illustrates an example in which a country-specific productivity innovation brings Home output above Foreign output, shifting the relative supply schedule from RS to RS'.

Figure 23.6: Relative demand and relative supply



The upward schedule in the figure (RD') illustrates the case in which, for a small enough ϕ, the wealth effects from TOT movements (captured by equation (23.25)) are the primary drivers of relative demand, dominating substitution effects. In response to a productivity innovation at home, relative home consumption and the demand for home goods both grow higher relative to their foreign counterpart, the home TOT strengthens, and the home RER appreciates,[30] resulting in a negative correlation of the exchange rate with both relative relative consumption and output. This suggests that models with incomplete markets (here in an extreme form), specified to reflect strong wealth effects of supply shocks, may provide an explanation of the Backus-Smith puzzle.

[30] Using equilibrium conditions, and further considering the limit of full home bias, $a_H \to 1$, we can show that $\mathcal{Q}_t = \frac{\hat{Y}_{H,t}-\hat{Y}^*_{F,t}}{2\phi-1}$.

The same applies in the bond economy. To see this, define real assets as $b_t \equiv B_t/P_t$, and the *ex post* (gross) real interest rate as $r_t \equiv (1/Q_t)(P_{t-1}/P_t)$, which in steady state is equal to β^{-1}. Then write the bond-economy wealth gap, up to first order around a symmetric steady state with zero net foreign assets (evaluated again at $\hat{Y}_{H,t} = \hat{D}_{H,t}, \hat{Y}^*_{F,t} = \hat{D}^*_{F,t}$), by substituting relative consumption in (23.23) with the relative budget constraint and imposing equilibrium:

$$\hat{\mathcal{W}}^{BE}_t = \sigma \left[\hat{D}_{H,t} - \hat{D}^*_{F,t} - 2(1-a_H)\widehat{\mathcal{T}}_t\right] - (2a_H - 1)\widehat{\mathcal{T}}_t + 2\sigma\beta^{-1}[\tilde{b}_{t-1} - \beta\tilde{b}_t]. \tag{23.27}$$

where $\tilde{b}_t = b_t - \bar{b}$ and $\bar{b} = 0$. This expression differs from the expression for $\hat{\mathcal{W}}^{FA}_t$ (23.25) (evaluated again at $\hat{Y}_{H,t} = \hat{D}_{H,t}, \hat{Y}^*_{F,t} = \hat{D}^*_{F,t}$) only by the last term in equation (23.27), which generally depends on both *current* and *future* relative prices. As agents smooth consumption in response to shocks, borrowing and lending may attenuate the wealth effects of inefficient relative price adjustment on current relative demand, compared to the financial autarky case. However, in the next section we show that intertemporal trade also magnifies the wealth effects of persistent or anticipated future productivity innovations on current demand. Remarkably, with these type of shocks, the condition on the trade elasticity for strong wealth and demand effects to materialize are different.

23.3.5 International transmission of productivity shocks via relative prices, wealth and demand

We now elaborate on the international transmission of supply shocks, focusing on their effects on relative prices and financial and capital flows. This will facilitate the comparison of the model with the empirical responses shown in Figure 23.5.

Terms of trade and real exchange rate. With complete markets, combining the relative demand equation, (23.24), with relative supply, and solving for the TOT yields:

$$\widehat{\mathcal{T}}_t = \frac{\sigma}{4a_H(1-a_H)(\sigma\phi - 1) + 1}(\hat{Y}_{H,t} - \hat{Y}^*_{H,t}). \tag{23.28}$$

In this expression, the denominator is unambiguously positive when $a_H \in (0.5, 1)$. So, in equilibrium, the home TOT unambiguously worsen and the RER depreciates with an increase in relative domestic productivity. Note that, other things equal, the coefficient on the productivity differential vanishes when goods are highly substitutable and is largest as ϕ approaches zero.

These results do not extend to cases with incomplete markets. Under financial autarky, using (23.25), the equilibrium TOT become:

$$\widehat{\mathcal{T}}_t = \frac{1}{1 - 2a_H(1-\phi)}(\hat{Y}_{H,t} - \hat{Y}^*_{H,t}). \tag{23.29}$$

In this expression, the denominator is not necessarily positive. The home TOT worsen following a domestic productivity improvement if and only if

$$\phi > \tilde{\phi}^{FA}_{TOT} \equiv 1 - \frac{1}{2a_H}.$$

For values of the elasticity below this threshold, a home productivity gain results in *stronger* TOT, as in the RD' schedule in Figure 23.6.

Comparing the expressions for the equilibrium TOT under complete markets, (23.28), and financial autarky, (23.29), yields a remarkable result. Observe that the equilibrium association between the TOT and relative productivity is identical across complete markets and financial autarky for $\sigma = \phi = 1$. This resonates with the contribution by Cole and Obstfeld (1991), where TOT movements move one-to-one with relative output, ensuring efficient sharing of productivity risk independently of trade in assets.[31] In this case, a relative rise in home output unambigously worsens the TOT under any financial market structure.

Cross-border real and financial flows. Key insights on the response of capital flows can be derived through further analysis of the complete-markets case. There are many equivalent ways to decentralize efficient state-contingent payments that support full risk sharing under complete markets. For our purposes, the most instructive approach consists of modeling them in analogy to movements in net foreign assets in the bond economy around the symmetric steady state with zero net foreign wealth. Following this approach, we introduce a 'notional real foreign asset,' denoted by $\widehat{\mathcal{B}}_t = \mathcal{B}_t - \overline{\mathcal{B}}$ with $\overline{\mathcal{B}} = 0$, defined as the cumulative real net exports scaled by steady-state output. Recall that real net exports are always uniquely defined in the complete-market allocation. Using the resource constraints together with the risk-sharing condition and equilibrium output, we then derive how net exports respond to productivity shocks, driving our notional flows and their accumulation over time, consistent with complete markets:

$$\widehat{\mathcal{B}}_t - \beta^{-1}\widehat{\mathcal{B}}_{t-1} = (1 - a_{\mathrm{H}})\,\sigma^{-1}\left[(2a_{\mathrm{H}}\,(\sigma\phi - 1) + 1 - \sigma)\,\widehat{\mathcal{T}}_t\right]. \tag{23.30}$$

The key result here is that net exports and their counterpart, financial flows, are proportional to the TOT (and hence the RER), but the relation can have either sign. To appreciate why, suppose that, at the initial steady state (i.e., $\widehat{\mathcal{B}}_{t-1} = 0$), the home TOT unexpectedly worsen because of a domestic output boom (i.e., $\widehat{\mathcal{T}}_t > 0$). Net exports turn positive and capital flows from the more productive country to the less productive country (hence from home into foreign) provided

$$\phi > \frac{1}{\sigma} + \frac{1}{2a_H}\left(1 - \frac{1}{\sigma}\right). \tag{23.31}$$

Note that this condition on the trade elasticities can be interpreted as the general-equilibrium counterpart of the classic Marshall-Lerner condition (on exports and imports elasticities) for an improvement in the trade balance in response to a worsening of the TOT.[32] Notably, with log utility (i.e., $\sigma = 1$), the condition will boil down to gross substitutability between home and foreign goods (i.e., $\phi > 1$).

[31] With $\phi = \sigma = 1$ and symmetric preferences in consumption, a country experiencing an increase in its (relative) output, will also experience a proportional fall in the international price of its goods. Hence, the value of national output remains constant, but consumption increases in both countries—residents abroad have higher income in real terms as they can buy the goods produced by the more productive country more cheaply.

[32] For a classic discussion of Marshall-Lerner conditions in modern general equilibrium models see Backus, Kehoe, and Kydland (1994a).

If the above condition is violated, financial resources will flow *into* the more productive country, which will thus run a trade deficit. It is worth spelling out the economics of this result, focusing again on the case $\sigma = 1$. In an efficient allocation, when home experiences an output boom, home households will increase their consumption of domestically-produced output. If $\phi < 1$ (home and foreign goods are gross complements), the home marginal utility from consuming home imports rises. It is then efficient for the foreign country to produce and export more. In a decentralized equilibrium of our economy with complete markets, this allocation is supported by a sharp deterioration in the home TOT, increasing the value of foreign output and triggering a flow of financial payments from abroad to home. In equilibrium, the home country runs an external deficit even though its output is high relative to foreign output.

In a bond economy, the general-equilibrium analog of the Marshall-Lerner conditions is more complex. For an elasticity sufficiently low, the model can generate a counter-cyclical trade deficit in response to productivity gains also when these gains lead to a TOT appreciation (hence irrespective of whether (23.31) holds).

23.4 The production economy

In the previous sections, we studied the transmission of productivity shocks in an endowment economy. Here, we extend our analysis to an environment in which labor supply is endogenous, so that we can discuss the ability of the model to account for the correlation of national business cycles seen in the data.

23.4.1 Model setup

The model is identical to the one in the previous section, with the following differences: (i) labor supply is endogenous and (ii) output is produced by competitive firms. In particular, the expected lifetime utility of the representative household is now:

$$U_t = \mathbb{E}_t \left[\sum_{j=0}^{\infty} \beta^j \left[u(C_{t+j}) - \zeta \nu(L_{t+j}) \right] \right], \tag{23.32}$$

where $\nu(L) = L^{1+\eta}/(1+\eta)$ is the instantaneous disutility from household labor supply (with $\eta > 0$ representing the inverse Frisch elasticity of labor supply), and $\zeta > 0$ is a constant parameter. With this specification, and regardless of the international financial market structure, the representative home household's labor supply in a competitive labor market will be determined by:

$$C_t^{-\sigma} w_t = \zeta L_t^{\eta}, \tag{23.33}$$

where w_t denotes the real wage. This expression equalizes the marginal benefit, in consumption terms, of an extra hour of work with its marginal disutility cost to the household. The preferences of the representative foreign consumer are defined analogously, with consumption C_t^*, labor L_t^*, expected lifetime utility U_t^*, as well as $\beta^* = \beta$, $\sigma^* = \sigma$, $\zeta^* = \zeta$ and $\eta^* = \eta$.

On the supply side, firms in each country produce specialized goods and are perfectly competitive. The representative home firm has the following production function:

$$Y_{H,t} = Z_t L_t^{\alpha}, \tag{23.34}$$

where Z_t represents an exogenous productivity process and $\alpha \in (0, 1]$ is the labor share. The representative foreign firm has an analogous production function, with foreign productivity Z_t^*. The stochastic processes for home and foreign productivity satisfies:

$$Z_t = \rho_1 Z_{t-1} + (1 - \rho_1)\overline{Z} + \varepsilon_t \quad \text{and} \quad Z_t^* = \rho_1 Z_{t-1}^* + (1 - \rho_1)\overline{Z} + \varepsilon_t^*, \tag{23.35}$$

where $\rho_1 \in (0, 1)$ is shock persistence and $\overline{Z} = 1$ denotes the steady-state value of productivity. In computational exercises, we impose that $\varepsilon_t, \varepsilon_t^* \sim \mathcal{N}(0, \sigma^*)$ where $\varepsilon_t \perp \varepsilon_t^*$ for all t.

The problem of the competitive firm is standard: hire labor to maximize profits, $p_{H,t}Y_{H,t} - W_t L_t$, with nominal wage W_t, such that real wages w_t equalize the marginal product of labor. Firms make zero profits in equilibrium (i.e., $p_{H,t}Y_{H,t} = W_t L_t$), so labor demand is determined by:

$$w_t = \frac{p_{H,t}}{P_t} Z_t L_t^{\alpha - 1}, \tag{23.36}$$

where $w_t \equiv W_t / P_t$. This expression highlights an important feature of the open-economy setting. The wage relevant for labor demand is not the real consumption wage w_t, but the nominal wage deflated by the product price—i.e., the product real wage $W_t / p_{H,t}$. Therefore, the relative price of domestic production in terms of consumption goods $p_{H,t}/P_t$ can act as a shifter of labor demand and thus introduces a channel of spillovers from shocks abroad.

23.4.2 Relative supply, relative demand, and global output

The relative demand is identical to the one in the endowment economy, given by equation (23.24). The supply of the home and foreign goods as a function of relative wealth and international relative prices can be derived by combining the optimizing conditions for labor demand (23.33) and labor supply (23.36), with the production function (23.34), for the two economies, using the wealth-gap definition (23.22) to substitute out relative consumption. Up to a first order, the home *relative to* foreign supply is:

$$\hat{Y}_{H,t}^S - \hat{Y}_{F,t}^{*S} = \frac{1 + \eta}{1 - \alpha + \eta}\left(\hat{Z}_t - \hat{Z}_t^*\right) - \frac{\alpha}{1 - \alpha + \eta}\left(\hat{\mathcal{W}}_t + \widehat{\mathcal{T}}_t\right), \tag{23.37}$$

where the superscript 'S' on output denotes supply and (recalling our definitions) η is the inverse Frisch elasticity, α captures the returns to scale in production, and Z_t^* is a home (foreign) productivity shock. The TOT now enter the relative supply expression (23.37) because the equilibrium demand for labor (23.36) depends on the relative price of the marginal product of labor in terms of domestic consumption, and hence varies with the relative price of imports.

Of course, the wealth gap also depends on relative prices, but in the case of complete market it is identically equal to zero, $\hat{\mathcal{W}}_t^{CM} = 0$. Since the expression multiplying $\widehat{\mathcal{T}}_t$ is always negative, (23.37) establishes that under complete markets the relative supply of home

in terms of foreign goods is invariably increasing in the relative price of the home goods as a function of parameters related to technology and labor supply. Under full risk sharing, a real appreciation is associated with a fall in relative consumption: via equation (23.33), a real appreciation is also associated with a rise in the relative supply of labor. Remarkably, under complete markets the slope of the relative supply curve does not depend on openness.

Under financial autarky, as already seen, the relative wealth gap is tightly linked to current real incomes and thus relative prices (see equation (23.25)). Substituting this into (23.37), the relative supply schedule becomes:

$$\hat{Y}^S_{H,t} - \hat{Y}^{*S}_{F,t} = \frac{1+\eta}{1-\alpha+\eta+\alpha\sigma}\left(\hat{Z}_t - \hat{Z}^*_t\right) + \frac{2\alpha(1-a_H)(\sigma-1)}{1-\alpha+\eta+\alpha\sigma}\widehat{\mathcal{T}}_t. \tag{23.38}$$

With no cross-border trade in assets, the relative supply of home in terms of foreign goods is either *increasing* or *decreasing* in the relative price of the home goods, depending on $\sigma \lessgtr 1$. When $\sigma > 1$, for any given amount of output supplied, the purchasing power of domestic residents rises with higher output prices, driving up leisure. That is, the wealth effects from a RER appreciation result in a fall in labor supply—the opposite of the complete-markets case. Remarkably, relative supply is independent of the TOT if preferences are logarithmic ($\sigma = 1$).[33]

To calculate global output, we combine the resource constraint with the sum of the national output supply, to express global output as a function of productivity shocks. Up to a first order, and dropping superscripts 'D' and 'S', we have the following log-linearized resource constraint:

$$\hat{Y}_{H,t} + \hat{Y}^*_{F,t} = \hat{C}_t + \hat{C}^*_t = \frac{1+\eta}{1-\alpha+\eta+\alpha\sigma}\left(\hat{Z}_t + \hat{Z}^*_t\right).$$

This shows that, globally, productivity gains unambiguously raise global output in equilibrium, regardless of the structure of international financial markets.

23.5 Substitution and wealth effects in the international transmission mechanism

Using the production economy presented above, we close our study with an educated selection of quantitative exercises meant to illustrate the workings of substitution and wealth effects in the international transmission of shocks. Figure 23.7 plots the effects of a positive home productivity shock under complete markets, financial autarky, and with riskless bonds only, for three alternative parameterizations. The first two columns bring together the results we have presented so far, demonstrating the effects of a transitory increase in home productivity for two values of the trade elasticity—where calibrations draw on the long-standing debates around the empirical estimation of this parameter.[34] The last column introduces a novel set

[33] See Heathcote and Perri (2002) for further analysis of the international real business cycle under financial autarky.

[34] Estimates of the elasticity at the macro level are typically low, while micro estimates high (see, e.g., Feenstra, Luck, Obstfeld, and Russ, 2018, and the references within). There are also differences across

of results, highlighting the specific, dynamic nature of the bond economy. Here, we assume that current shocks create expectations of further increases in output in the future, impacting wealth and demand as national agents reassess the present discounted value of their current and future income and smooth consumption by borrowing and lending internationally.

Strong substitution effects. In the first two columns, we plot impulse responses to a home productivity shock, which we model using the AR(1) process in (23.35) with persistence $\rho_1 = 0.97$. In column (a) we set $\phi = 1.5$. As shown analytically above, in this case, substitution effects dominate regardless of the degree of cross-border risk sharing, such that the home TOT deteriorates in response to the shock and cross-border spillovers are negative.

Under complete markets, the correlation between home and foreign business cycles is counter-factually (and strongly) negative. Substitution effects from the home RER depreciation boost home production relative foreign—home consumption is less volatile than output. Output co-movements remain negative also under financial autarky, although moderate in comparison: home output rises by less—and foreign output falls by much less—than under full risk sharing. The depreciation of the home TOT now erodes the purchasing power of home consumers, while boosting that of foreign consumers.[35] Remarkably, impulse responses for the bond economy lie between the other two cases. The bond economy is quantitatively different from the complete-markets benchmark, as intertemporal consumption smoothing falls short of replicating full risk sharing. Yet capital flows in the same direction. In both cases (holding (23.31) and its counterpart for the bond economy), home real net exports are positive, such that resources flow from the home economy to foreign.

Strong wealth effects with a low trade elasticity. In column (b), we plot impulse responses to the same AR(1) home productivity shock, but lower the trade elasticity to $\phi = 0.3$. In this case, business cycles are positively correlated and under, imperfect risk sharing, RER tend to appreciate with the home relative output boom—replicating qualitatively the empirical responses shown in Figure 23.5.

Under complete markets, the TOT deteriorates by much more than in the previous case, such that home residents are poorer. Remarkably, the notional financial flows run from foreign to home, in spite of the fact that home experiences a rise in productivity. With incomplete markets, wealth effects dominate, driving the response of the equilibrium TOT in the opposite direction. The relative price of home output improves strongly in financial autarky and the bond economy. With incomplete risk sharing, foreign output is higher relative to the complete-market case—the home appreciation boosts demand by home residents, who increase consumption of both domestic and the foreign goods. However, consumption is now counterfactually more volatile the output. Note that the home country runs a persistent trade deficit both under complete markets and in the bond economy, even though the TOT move in opposite directions.

horizons, with elasticities typically higher in the long run than in the short run (Boehm, Levchenko, and Pandalai-Nayar, 2023). Example of models with short vs. long-run elasticities include those with distributive capital (Crucini and Davis, 2016), or its analog in terms of 'customer list' capital (Drozd and Nosal, 2012), as well as models with nested CES (Cooley and Quadrini, 2003).

[35] Recall from equation (23.21) that there is a linear relationship between TOT and RER in our approximation.

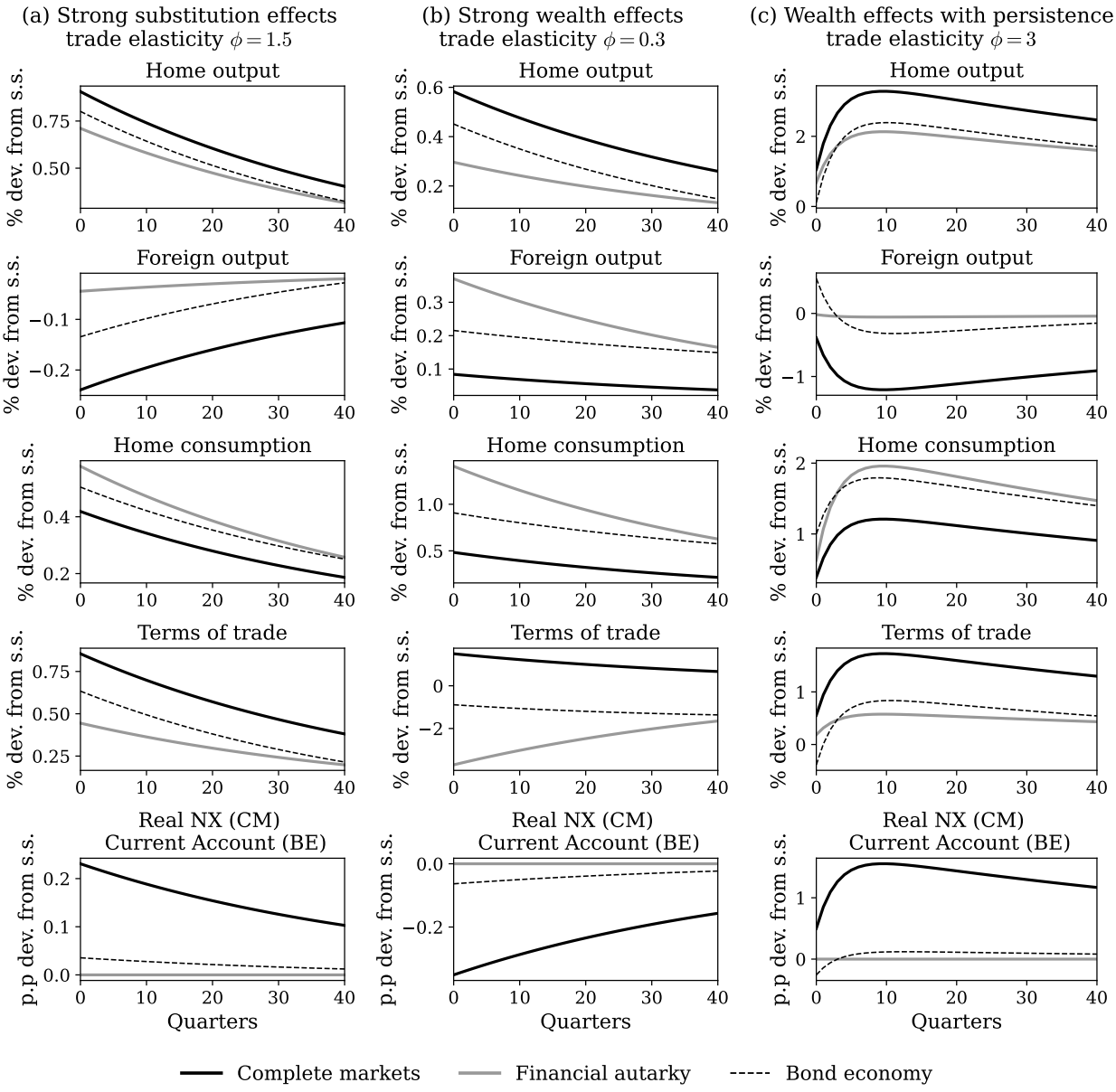


Figure 23.7: International transmission of positive home productivity shock

Note: impulse response to positive and persistent 1% increase in home productivity. In all columns: $\beta = 0.99$ (quarterly), $\sigma = 2$, $\eta = 1$, and ζ such that steady-state labor supply is $1/3$, $a_H = 0.7$, $\alpha = 1$ $\overline{Z} = 1$. Bond economy made stationary using Uzawa-style preferences. *Persistence*: In columns (a) and (b) home productivity is an AR(1) process with persistence $\rho_1 = 0.97$; in column (c) home productivity follows an AR(2) process with persistence $\rho_1 = 1.7$ and $\rho_2 = -0.703$. *Trade elasticity*: in column (a) $\phi = 1.5$; (b) $\phi = 0.3$; and (c) $\phi = 3$. Parameterization is symmetric across countries.

Strong wealth effects with high shock persistence and high trade elasticities. In column (c), we bring forward a different way to model cross-border wealth effects of shocks. First, we assume a hump-shaped AR(2) productivity process (with persistence parameters ρ_1 and ρ_2) so that, in effect, an increase in home productivity today also conveys news about even *higher* future productivity. It is worth noting that this is a convenient reduced-form way to capture output dynamics in the model without capital accumulation.[36] Second, we assume that home and foreign goods are highly substitutable. Under this parameterization, the expected real income from higher home output in the future is positive, i.e., higher quantities are not associated with a more than proportional fall in output prices. This means that, under incomplete markets, the shocks have a significant impact on home wealth.

The exercise highlights macroeconomic dynamics specific to the bond economy. Under complete markets and financial autarky, substitution effects dominate in equilibrium: higher home productivity is associated with a RER depreciation, but business cycles are counterfactually negatively correlated.[37]

The dynamics of the bond economy overturn these negative cross-border spillovers and ensure that consumption is less volatile than output—qualitatively mirroring the empirical responses shown in Figure 23.5. Faced with the prospect of even higher productivity in the future, home households have an incentive to run a current-account deficit and borrow to raise current consumption. Consumption smoothing thus dampens the home TOT deterioration relative to complete markets and financial autarky—it even improves them on impact, replicating the appreciation seen in Figure 23.5. Facing these changes in relative prices, foreign households produce more in the first periods after the shock. As a result, cross-border transmission is positive. Over time, the home terms of trade eventually deteriorates, and the cross-border transmission turns negative as home runs trade surpluses to repay its debt.[38]

Taking stock. Taken together, these results suggest that, to bring the benchmark model closer to the empirical responses, an increase in output has to be associated with a 'boom' in demand. Conditional on productivity shocks, the model features that are crucial for generating these facts are incomplete financial markets and either a low trade elasticity (Figure 23.7, column (b)) or shock processes that result in a persistent rise in the growth rate of output (Figure 23.7, column (c))—see Corsetti et al. (2008b) for a generalization of this analysis to models with capital accumulation.

[36] A hump-shaped output profile mimics the dynamics obtained from a model with capital accumulation, as investment responds steadily to persistent productivity gains (see, e.g., Corsetti et al., 2008b). See early work by Baxter and Crucini (1995), assessing the implications of increasing the persistence of productivity shocks.

[37] Observe that in our model without capital, anticipations of future output do not affect the current allocation under both financial autarky and complete markets. Under financial autarky, agents have no instruments to smooth consumption intertemporally. Under complete markets, future productivity variations are completely insured by financial contracts at the time they materialize. Instead, anticipation of future output dynamics becomes crucial in the bond economy.

[38] Strong wealth effects can also be obtained in model assuming that the business cycle is driven by stochastic shifts in the growth rate (Aguiar and Gopinath, 2007), 'news' shocks (Jaimovich and Rebelo, 2008) or input adjustment costs and cointegrated productivity shocks across countries (Rabanal and Rubio-Ramírez, 2015).

23.6 Richer frameworks

The framework developed and studied in this chapter provides a core building block for modern open-economy macroeconomics. Armington aggregators of domestic and foreign goods (and/or intermediate) inputs are a key feature of seminal contributions to the international real business cycle literature (e.g., Backus et al., 1992; Backus, Kehoe, and Kydland, 1994b). Building on this same structure, the New Keynesian open-economy model——as well as earlier contributions to the New Open Economy Macroeconomics (NOEM) literature (e.g., Obstfeld and Rogoff, 1995; Corsetti and Pesenti, 2001; Benigno and Benigno, 2003; Devereux and Engel, 2003)—departs from the assumption of perfect competition, instead assuming that firms operate under monopolistic competition with nominal rigidities.[39]

Richer specifications make the model better suitable to confront the data. Investment and capital accumulation enriches the dynamic of the model, helping generating countercyclical current account balances: a positive productivity shock to domestic technology simultaneously raises output and the demand for imported capital goods—see the literature building on Backus et al. (1994b), and especially the discussion in Raffo (2008). Modeling a non-tradable good sector in each economy introduces a new relative price (nontradable to tradables) that breaks the proportionality between real exchange rate and terms of trade, potentially magnifying currency and misalignment volatility and the wealth effects from shocks (see Stockman and Tesar, 1995; Corsetti et al., 2008a, 2014). In addition to fostering data-rich micro-to-macro research, modeling firm dynamics, distributive trade and global supply chains contribute to account for possibly state-contingent and time-varying trade elasticities, driving destination-specific price and markup adjustment by exporters and possibly contributing to explain the cross-border correlation of economic activity (see, e.g., Ghironi and Melitz, 2005; Atkeson and Burstein, 2008; Johnson, 2014; Bergin and Corsetti, 2020).[40] Complementing work on destination-specific price adjustment, introducing nominal frictions can help to explain why import prices remain stable in local currency despite a high volatility of nominal exchange rates, potentially paving the way to studies addressing the Mussa Puzzle. More generally, monetary models featuring price or wage stickiness make it possible to study the international transmission of nominal demand shocks, and explore normative implications for the optimal design of stabilization policy in global settings (see Corsetti, Dedola, and Leduc, 2010, 2023, for further exposition).

From a global perspective, studies into financial spillovers highlight important asymmetries in the global transmission associated with the dominant position of U.S. and the U.S. dollar in the international financial system (Gourinchas and Rey, 2007; Miranda-Agrippino and Rey, 2020). Explicitly modeling cross-border financial intermediation subject to frictions can capture these asymmetries, as shown by Gabaix and Maggiori (2015). Building on the approach by these authors, recent literature has shown that a rich specification of the

[39] Numerous recent studies have focused on nominal rigidities in a world with dominant currency pricing (e.g., Gopinath, Boz, Casas, Díez, Gourinchas, and Plagborg-Møller, 2020; Egorov and Mukhin, 2023; McLeay and Tenreyro, 2025).

[40] Modeling trade costs along manufacturing or distribution chains introduces a wedge between the (market) trade elasticity of imports, relevant for firms' decisions, and the elasticity in production (Corsetti and Dedola, 2005). Corsetti, D'Aguanno, Dogan, Lloyd, and Sajedi (2023) embeds global value chains into the model outlined in this chapter.

model can account for international business cycle facts jointly with international finance facts. The latter includes the apparent 'disconnect' of exchange rates from fundamentals, the close correlation of the real and nominal exchange rate (the Mussa puzzle), as well as the Uncovered and Covered Interest Parity Puzzles (Itskhoki and Mukhin, 2021a,b). Financial market imperfections implying imperfect risk sharing are essential for these results. Literature has specifically stressed two imperfections. The first is moral hazard, giving rise to financial constraints that limit the risk-bearing capacity of intermediaries and makes the recourse to intermediation to borrow and lend internationally costly for the households.[41] The second is noise trading, that inject exogenous volatility in the currency and asset markets (see, e.g., Devereux and Engel, 2002; Jeanne and Rose, 2002).[42]

Relative to the rich and growing body of open-macro literature, this chapter has shown that the international transmission mechanism is best understood modeling frictions in the financial and good markets that undermine efficient risk sharing and create an economically meaningful feedback between equilibrium international prices and current and anticipated future domestic output on the one hand, and cross-border wealth and demand on the other. While following this approach may address some of the "puzzles" in the literature, others remain outstanding. The field is now moving further towards integrating trade and macro, allowing for heterogeneity across households as well as firms/sectors, and rethinking global equilibria in the context of rising geopolitical fragmentation and strategic policy games undermining cross-border policy cooperation.

[41]Because of this cost, dynamically, the model shares the same properties of the bond economy with costly cross-border bond holdings as discussed by Schmitt-Grohé and Uribe (2003). Yet, explicitly modeling balance sheets creates opportunities to explore the role of international reserve policy, swap lines and other forms of international liquidity interventions (Bahaj and Reis, 2022).

[42]Studies into the structure of international financial market, where information, incentives and/or exogenous shocks may drive portfolio reshuffling unrelated to macro fundamentals, include Bacchetta and Van Wincoop (2006) and Bippus, Lloyd, and Ostry (2023).

Chapter 24

Sovereign debt and default risk

Juan C. Hatchondo and Leonardo Martinez

24.1 Introduction

The bulk of this textbook assumes agents have full commitment to repay their debt. In this chapter, we remove that full-commitment assumption. An immediate implication is that the debt sustainability calculations from Chapter 15—which were based on constant risk-free rates—no longer hold. When repayment is uncertain, interest rates become endogenous, rising with the perceived probability of default. Sustainability assessments must therefore account not only for debt levels and fiscal policy, but also for the incentives and constraints that shape repayment decisions.

Historically, sovereign defaults have been most closely associated with emerging markets, most notably the wave of debt crises across Latin America in the 1980s, when a combination of high global interest rates, collapsing commodity prices, and excessive borrowing triggered widespread defaults. However, sovereign risk is by no means confined to developing economies. The European debt crisis of 2011–2012 revealed similar vulnerabilities in advanced economies, as Greece, Portugal, Spain, and Ireland faced surging borrowing costs and questions about their solvency within the Euro area. These episodes highlight that the underlying mechanisms we study—how rising debt and default risk interact to constrain borrowing and raise interest rates—operate in economies of heterogeneous size and income levels, from emerging markets to advanced economies.

A further motivation for studying sovereign default comes from a distinctive set of business cycle patterns, particularly evident in emerging markets. As documented in Chapter 23, these economies exhibit greater output volatility and, more strikingly, consumption that is more volatile than output—contrary to the standard consumption-smoothing logic. This excess volatility is closely tied to their external borrowing behavior, which tends to be strongly procyclical: countries borrow heavily in booms and repay during recessions. One leading explanation is that credit access shrinks in downturns.

These patterns point directly to the mechanisms at the heart of our analysis. To understand them, we need a framework in which sovereign risk shapes borrowing opportunities and, in turn, macroeconomic volatility. This chapter develops such a model. We begin by documenting key features of business cycles and sovereign debt markets across countries. We then introduce a canonical sovereign-default model in which a benevolent government bor-

rows abroad but can choose to default. The model captures the central tension between the desire to smooth consumption over time and the temptation to renege on debt obligations. Unlike the frictionless models in Chapters 23 and 15, we explicitly incorporate endogenous default risk.

24.2 Empirical patterns

We begin by documenting key features of business cycles and sovereign debt markets across countries: the excess volatility of consumption, the cyclical behavior of sovereign spreads, the frequency and consequences of default episodes, and the patterns of "debt intolerance."

24.2.1 Excess volatility of consumption

Aguiar and Gopinath (2007) and Neumeyer and Perri (2005) notice key differences in the dynamics of aggregate variables in advanced vs. emerging-market economies. They show that while aggregate consumption is less volatile than income in small open advanced economies, it is more volatile than income in emerging economies. The extra consumption volatility is at odds with the standard consumption-smoothing motive, which prescribes that households save a fraction of the extra income earned during economic booms to buffer the consumption drop during recessions. As shown in Table 23.1 in Chapter 23, this pattern remains evident in more recent data: the consumption-to-output volatility ratio is higher in emerging markets than in advanced economies, and the overall volatility of output is also significantly larger. Although the differences are somewhat less pronounced than those documented by Aguiar and Gopinath (2007), they persist and continue to motivate the theoretical frameworks developed in this chapter.

The extra consumption volatility in emerging economies is accounted for by their external borrowing behavior. In an open economy, agents can dissave by liquidating capital or by accumulating net external liabilities. As shown by Table 23.1, there are no significant differences in investment behavior between advanced and emerging economies, implying that most of the extra consumption volatility in emerging economies is accounted for by their external borrowing behavior. This is reflected in a more countercyclical trade balance for emerging economies. A typical pattern in several emerging economies is to borrow from the rest of the world to finance trade balance deficits during booms, and partially repay their external liabilities by running trade balance surpluses during recessions. Again, this is at odds with standard consumption smoothing.

What accounts for the apparently puzzling borrowing behavior of emerging economies? Alvarez-Parra, Brandao-Marques, and Toledo (2013), García-Cicco, Pancrazi, and Uribe (2010), and Neumeyer and Perri (2005) argue that borrowing opportunities play an important role. They present different versions of the small open economy model expanded with various shocks and the following (simplified) aggregate budget constraint:

$$C_{t+1} + K_{t+1} = Y_t + K_t(1-d) + q(Y_t, \eta_t, B_{t+1})B_{t+1} - B_t. \tag{24.1}$$

Here, C denotes aggregate consumption, Y the level of output, K the stock of capital with depreciation rate d, B denotes the net external liabilities, q denotes the price at which

the economy can issue debt, and η denotes a shock to the price q.[1] Alvarez-Parra et al. (2013), García-Cicco et al. (2010), and Neumeyer and Perri (2005) find that a price q that increases with income and decreases with debt plays a key role in accounting for the strong countercylical trade balance, and thus, procyclical external borrowing.

24.2.2 Sovereign defaults

Both the academic and policy literature on sovereign defaults typically use the definition of a default event proposed by credit-rating agencies. This definition identifies as a default event every episode in which the sovereign makes a debt restructuring offer with terms that are less favorable to creditors than the original debt terms. Thus, default events include both legal defaults in which the sovereign breaches the original debt contract, and "pre-emptive" debt restructurings (renegotiations of debt terms before a payment has been missed). Asonuma and Trebesch (2016) find that 38% of debt restructurings between 1978 and 2010 were pre-emptive.

The most frequently used measure of creditors' losses after a debt restructuring is the present-value "haircut," which is defined as

$$Haircut = 1 - \frac{\sum_{t=1}^{\infty} \frac{x_t^{Post}}{(1+i)^t}}{\sum_{t=1}^{\infty} \frac{x_t^{Pre}}{(1+i)^t}}, \tag{24.2}$$

where x_t^{Post} denotes the post-restructuring debt payment obligations in period t, x_t^{Pre} denotes the pre-restructuring debt payment obligations in period t, and i denotes the bond yield prevailing immediately after the restructuring (see, e.g., Sturzenegger and Zettelmeyer, 2005). Cruces and Trebesch (2013) calculate that the average present-value haircut in sovereign debt restructurings is 37%.

Sovereign defaults have been a relatively frequent event in emerging economies. The updated dataset of Asonuma and Trebesch (2016) and Asonuma, Niepelt, and Ranciere (2017) records 201 sovereign defaults to private creditors between 1970 and 2020. Figure 24.1 shows that, at the end of 2022, emerging economies' sovereign debt in default represented 2% of their GDP.

Sovereign default facts

Aguiar and Amador (2014) summarize the key characteristics of sovereign defaults. First, they tend to occur in bad times. Tomz and Wright (2007) find that more than 60% of sovereign defaults occurred in years when the GDP was below trend. This is consistent with the aforementioned countercyclicality of the borrowing cost in emerging economies. Secondly, sovereign defaults are followed by years of negotiations between the sovereign and its creditors. Thirdly, defaults end when sovereigns settle their debt in default with a debt restructuring (in which defaulted debt instruments are exchanged

[1] If the economy saves (it chooses $B_{t+1} < 0$) by buying foreign bonds that pay the global risk-free interest rate r, the bond price is $q = 1/(1+r)$.

for new debt instruments). Asonuma and Trebesch (2016) and Asonuma et al. (2017) report an average of 3.2 years between the onset of default and its settlement with creditors, with a significant heterogeneity in default durations and haircuts. For example, preemptive restructurings have an average negotiation time of 12 months, and post-default restructurings have an average negotiation time of 60 months.

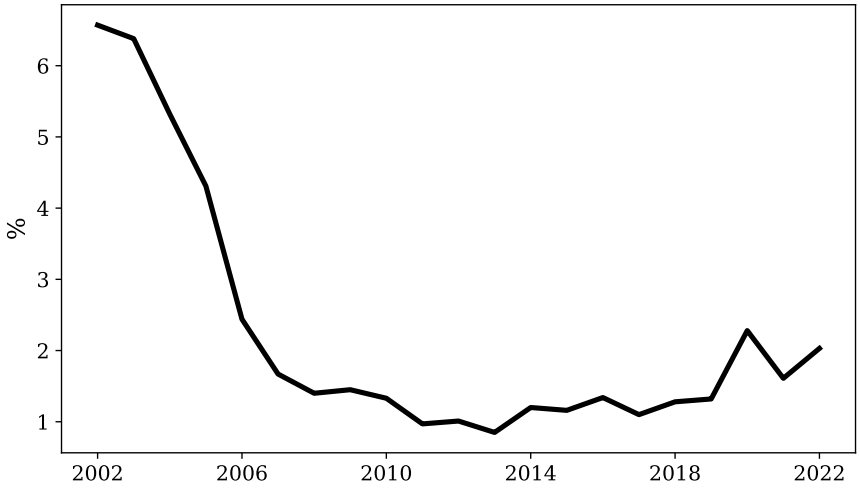


Figure 24.1: Emerging economies' public debt in default.

Notes: The series is expressed as a percentage of the combined GDP of emerging economies. Source: 2023 Bank of Canada Sovereign Default Database.

24.2.3 Sovereign spreads

As a consequence of default risk, sovereign bond yields command a premium over safer bonds.[2] The Emerging Markets Bond Index (EMBI) computed by J.P. Morgan is the most commonly used reference for bond yields for sovereign debt in emerging economies. The index measures the sovereign yield spread, i.e., the extra yield of U.S. dollar-denominated bonds over similar bonds issued by the U.S. government.

Table 24.1 shows that sovereigns in emerging economies pay a significant and volatile premium when they borrow from international markets. Also, the fourth column shows that the premium increases when the GDP is below trend. The high, volatile, and countercyclical sovereign spreads here are consistent with the findings in Alvarez-Parra et al. (2013), García-Cicco et al. (2010), and Neumeyer and Perri (2005). The last section of this chapter presents a quantitative sovereign default model in which the spread is endogenous and displays the same behavior as in the data. We show that this feature plays an important role in accounting for the distinctive business cycle fluctuations of emerging economies.

Table 24.2 illustrates the potential determinants of the sovereign spread. It presents the

[2]The bond yield at time t represents the hypothetical return an investor would make if it holds the bond until the maturity date and the bond is not defaulted on.

Table 24.1: Sovereign spread yield

	E(Spread)	σ(Spread)	ρ(Spread, GDP)
Argentina	732	377	-0.5
Brazil	514	380	-0.2
Ecuador	1021	647	-0.6
Korea	162	116	-0.7
Malaysia	174	115	-0.3
Mexico	318	221	-0.4
Peru	307	199	-0.1
Philippines	283	172	-0.3
Slovak Republic	57	36	-0.6
South Africa	242	111	-0.3
Thailand	155	112	-0.5
Turkey	385	199	-0.5
Average Emerging	362	224	-0.4

Note: The sovereign spread is expressed in basis points. The spread series is quarterly and spans the period Q1 1994 - Q4 2019 whenever data is available. We removed periods in which the sovereign is in default. The series for aggregate GDP was logged and filtered using the Hodrick-Prescott filter with a smoothing parameter of 1600.

results of a fixed-effects panel regression that estimates

$$\log\left(\text{Spread}\right)_{it} = \alpha + \beta X_{it} + \delta_i + \eta_{it}, \qquad i = 1, ..., N, \quad t = 1, ..., T,$$

where i denotes the country index, t denotes the year, X_{it} is a vector of control variables for country-specific and global macroeconomic factors, δ_i are country fixed effects; and η_{it} represents disturbances that are independent across countries and time. The table shows that the spread (i) increases with the total government debt, (ii) decreases with GDP growth, (iii) decreases with foreign reserves owned by the sovereign, and (iv) increases with the VIX index (reflecting global financial conditions).[3] While we cannot extract causal relationships from Table 24.2, the findings suggest a strong link between the sovereign spread, sovereign borrowing, domestic shocks to economic activity, sovereign asset accumulation, and global financial shocks.

Costs of sovereign risk. On top of the adverse effects of sovereign risk on economic stability and business cycles discussed above, following the Eurozone crisis, a recent branch of the literature has focused on the adverse effects of sovereign risk on investment and growth. For example, Arellano, Bocola, and Bai (2024) find that a 100 basis points increase of the sovereign spread leads to a 64 basis points increase in firms' borrowing cost. They also calculate that in 2012, if the sovereign spread had not increased, real GDP in Italy would have fallen 3.2 percent instead of 6.4 percent.

[3]The results in the table are similar to the ones reported in Jaramillo and Tejada (2011), Akitoby and Stratmann (2008), and other studies.

Table 24.2: Panel regressions

	Coeff.	SE
Public debt to GDP	0.020	(0.004)
Real GDP growth	-0.042	(0.006)
Reserves to GDP	-0.033	(0.017)
Net gov borrowing to GDP	0.020	(0.014)
VIX	0.034	(0.003)
Observations	523	
R-squared	0.77	
Number of countries	33	

Note: We use annual data from a sample of 33 emerging market countries spanning from 1994 to 2018. Robust standard errors are in parentheses.

The costs of sovereign risk were also apparent during COVID-19, when countries with higher risk suffered a stronger deterioration of their borrowing opportunities and thus had more limited resources to mitigate the adverse effects of the pandemic. Figure 24.2 shows that the spread increase after COVID-19 was typically larger for countries that were already facing a higher sovereign spread (reflecting higher risk) before the pandemic. In particular, countries with a pre-COVID spread below 300 basis points only suffer a spread increase below 100 basis points.

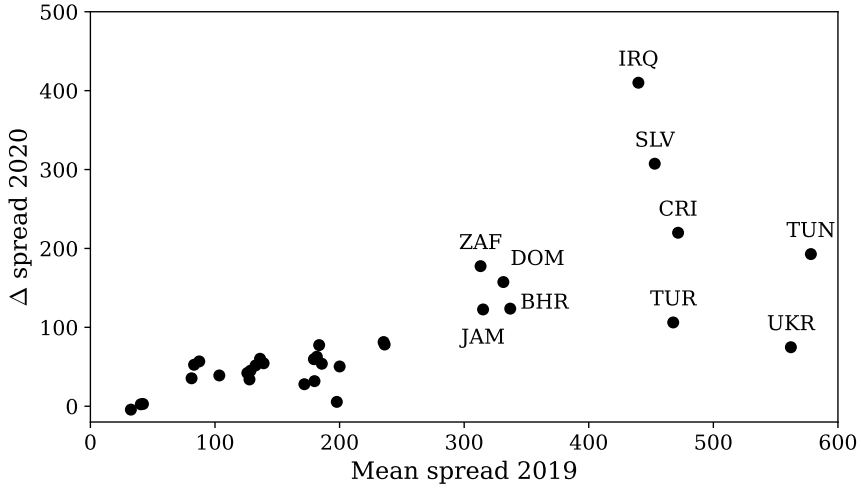


Figure 24.2: EMBI Spread increase after COVID-19.

Notes: The (average daily) spread and spread change (2020 average minus 2019 average) are expressed in basis points. We removed countries with a spread higher than 1000 basis points in 2019 (Argentina, Lebanon, and Venezuela). Source: Bloomberg.

24.2.4 Debt intolerance

The inability of sovereigns to commit to repaying their debt constrains their borrowing capacity. This feature has been emphasized in the literature on debt intolerance. Debt intolerance can be understood as the mapping from public debt to sovereign default risk. Debt intolerance varies both across countries and over time (Reinhart, Reinhart, and Rogoff, 2015; Reinhart, Rogoff, and Savastano, 2003). Figure 24.3 illustrates variations in the mapping between public debt ratios and sovereign default risk (measured using credit default swap spreads). First, the figure shows that there is significant heterogeneity in the debt-spread mapping. Economies like Thailand or Uruguay can feature public debt ratios above 50% while incurring modest default risk, while Turkey incurs a substantial default risk at lower debt ratios. Second, some economies accumulated significant debt without increasing their default risk (e.g., China and Thailand), while others experienced an increase in default risk after debt accumulation (e.g., Brazil). Finally, the mapping from debt and spread significantly shifts over time. For instance, default risk in Portugal, Greece, and Latvia declined significantly between 2010 and 2023, without substantial debt reductions (note also that before 2008, markets were pricing a nearly zero default risk for most Eurozone economies).

We will show how assuming shocks to global risk premia and/or aggregate income, sovereign default models can account for the volatility of default risk observed in the data for one country. However, those shocks alone are insufficient in accounting for the cross-country heterogeneity in the debt-spread mapping and some of the time variation in debt levels within countries. This suggests the presence of structural heterogeneity in default incentives across countries and time.

The difficulty for governments in emerging economies of issuing debt in their own currency has been associated with the debt intolerance problem in these economies. This contrasts with the landscape in advanced economies, which typically issue debt in their own currencies. Eichengreen and Hausmann (1999) refer to the issuance of debt denominated in foreign currency as the "original sin." Issuing debt in foreign currency is problematic in part because the exchange rate tends to depreciate in bad times, increasing the foreign-currency debt burden (in comparison with the government's income, which is mostly determined in local currency; Hausmann, 2003).

The left panel of Figure 24.4 shows that the share of sovereign debt issued in foreign currency is positively correlated with the sovereign spread. The difficulty emerging economies have had in issuing debt in their own currency has been linked to pitfalls in their ability to conduct monetary policy (Du, Pflueger, and Schreger, 2020; Engel and Park, 2022; Ottonello and Perez, 2019). Some countries are increasingly overcoming this problem: Du and Schreger (2016) show how sovereigns in emerging economies are borrowing more in their local currency. The right panel of Figure 24.4 shows that many emerging economies have managed to issue less foreign currency sovereign debt.

24.2.5 Remedies

The most notable difference between sovereign debt and household or corporate debt is that the former lacks a well-specified bankruptcy protocol. One avenue that sovereigns in emerging economies have followed to enhance their repayment credibility is to issue debt in

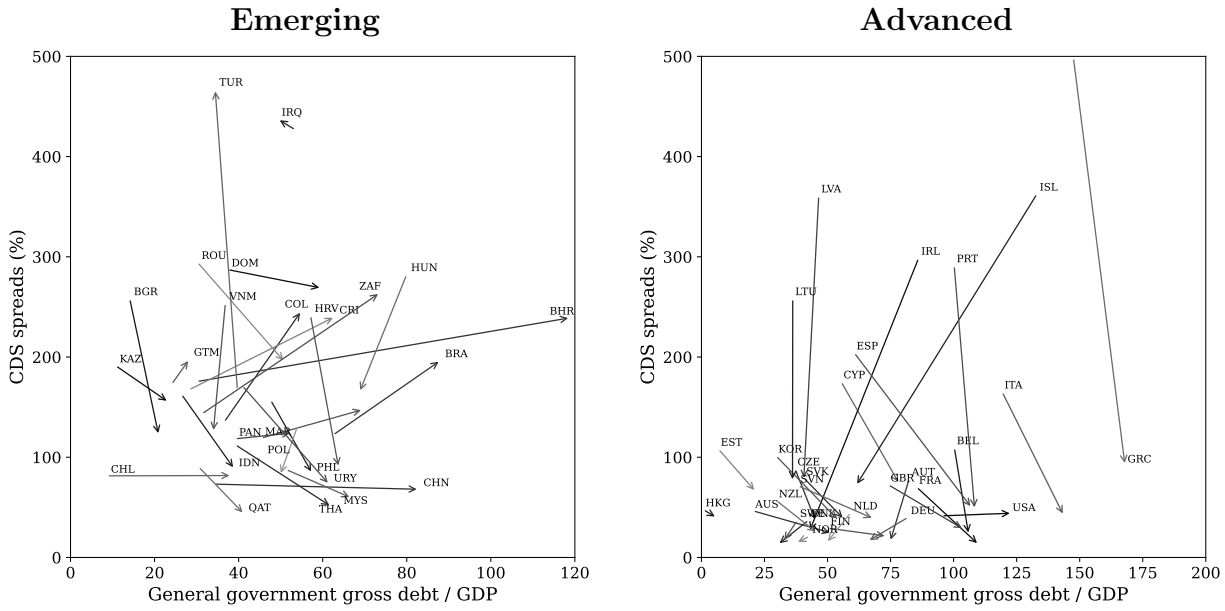


Figure 24.3: Credit default swaps (CDS) spread and public debt ratios for emerging and advanced economies.

Notes: The origin point of each arrow illustrates the debt-spread combination in 2010, and its end point the debt-spread combination in 2023. We use the CDS spread because it enables us to expand the sample of countries to advanced economies. Sources: Bloomberg and IMF-WEO database.

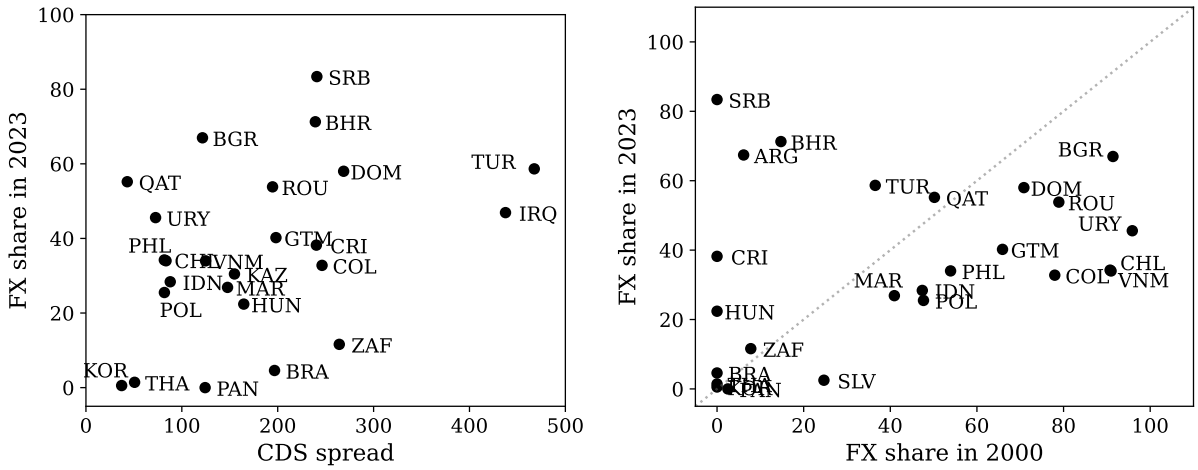


Figure 24.4: Spreads and foreign currency sovereign debt.

Notes: The left panel shows a positive association between the CDS spread (in basis points) and the share of foreign currency sovereign debt in emerging economies. The right panel illustrates how the share of foreign currency sovereign debt in emerging economies changed from 2000 to 2023. Source: WEO database and Bloomberg.

foreign financial centers, and thus under the jurisdictions of foreign courts.[4] However, in

[4]Sovereigns also increase their cost of defaulting by introducing debt covenants that include acceleration and cross-default clauses. Acceleration clauses allow creditors to call all future payments in a bond in case

contrast to the ability of courts to force payments by private agents, it is very difficult for foreign courts to enforce payments by sovereigns. Thus, while posting collateral is a strategy often followed by households and firms to obtain loans at a lower interest rate, with a few notable exceptions, sovereigns have a limited ability to credibly pledge collateral. This is because the principle of sovereign immunity limits creditors' right to confiscate sovereign assets.

Fiscal rules as remedies to sovereign risk. Figure 24.5 shows that an increasing number of countries are adopting fiscal rules, in part to limit "excessive" borrowing, sovereign risk, and costly sovereign default episodes.[5] The figure also shows that the bulk of countries adopting fiscal rules are limiting the debt level (limits to the fiscal budget balance or the fiscal expense also constrain the debt level). Aguiar, Amador, and Fourakis (2020) and Hatchondo, Martinez, and Roch (2022) use a sovereign default model to quantify the benefits of debt limits.

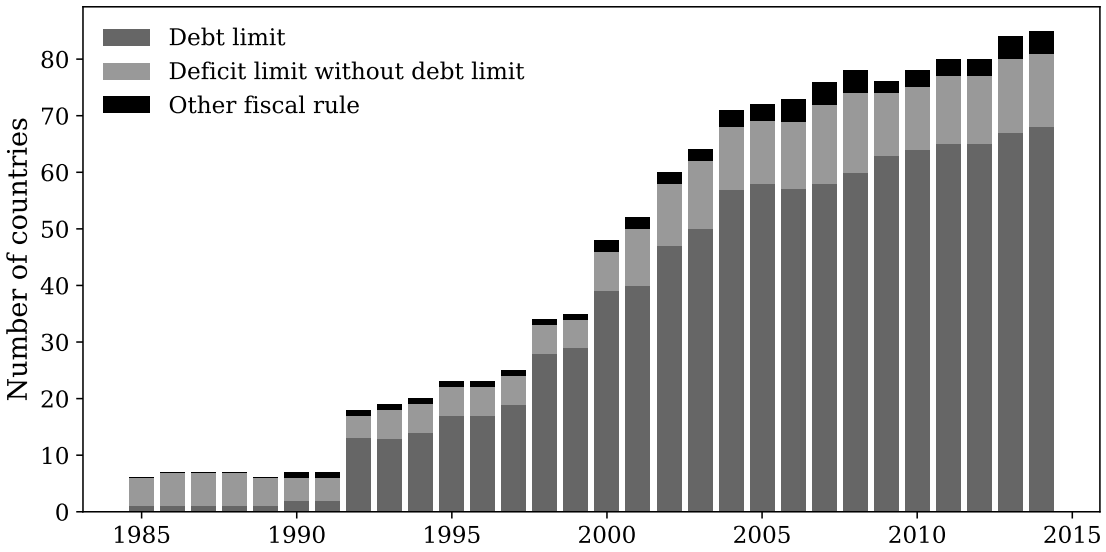


Figure 24.5: Number of countries with fiscal rules.

Source: IMF Fiscal Rules dataset.

However, the variation in debt intolerance across countries and time described above presents challenges for the use of debt levels as anchors in fiscal rules. For example, the variation of debt intolerance across countries makes it challenging to introduce the same debt limit to a group of heterogeneous countries, as with the Maastricht treaty for Eurozone members. The common limit is likely to excessively restrain public borrowing for some of the countries in the group with low debt intolerance (that could otherwise expand debt without significant increases in default risk). At the same time, the common limit may be

the government defaults on one payment of that bond. Cross-default clauses state that a default in any government debt obligation constitutes a default in the contract containing the cross-default clause.

[5]Fiscal rules in Figure 24.5 impose at least one and often more than one numerical target. These targets may limit the level of debt, the budget balance, revenues, and expenditures.

ineffective for economies with high debt intolerance (that could display significant sovereign risk for lower debt levels). Likewise, a time-invariant fiscal rule can become inadequate for economies that present time-varying debt intolerance.

The limitations of debt levels as fiscal anchors were underscored in recent debates about reforming the European Union fiscal framework. Blanchard, Leandro, and Zettelmeyer (2020) propose to replace debt-limit fiscal rules with fiscal standards (i.e., qualitative prescriptions that leave room for judgment) guided by probabilistic analysis that reflects the level of sovereign risk (as done by the IMF for their sovereign risk assessment for market-access countries; IMF, 2022). Similarly, Hatchondo et al. (2022) propose to use the sovereign spread (reflecting sovereign risk) as the anchor for fiscal rules. Furman and Summers (2020) suggest focusing on the government's interest payments to account for fluctuations in the (risk-free) interest rate.

Low-income countries

One key characteristic of low-income countries is the predominant role of official debt, often in the form of concessional loans (e.g., with an interest rate below the market rate or with grace periods). While official debt is sometimes restructured, the majority of official debt restructurings do not feature face-value write-offs (Cheng, Diaz-Cassou, and Erce, 2016; IMF, 2023), with some well-known exceptions such as the 1996 Heavily Indebted Poor Countries Initiative. For example, the Debt Service Suspension Initiative for the official bilateral debt of poorer countries implemented during COVID-19 (Lang, Mihalyi, and Presbitero, 2023) did not include face-value write-offs. Thus, in contrast with the sovereign spread for emerging economies discussed above, the interest rate governments pay for official debt typically does not reflect default risk. The preponderance of official debt can have a significant effect on sovereign risk, which is the focus of the sovereign default literature as well as this chapter. In fact, the ability of governments to obtain debt in concessional terms is the key criterion of the International Monetary Fund for deciding on the framework used for assessing the risk of debt distress in a country (IMF, 2013). Horn, Reinhart, and Trebesch (2021) discuss the increasing importance of China as a lender. This includes the "Belt and Road" initiative to finance and build infrastructure.

The rest of the chapter presents first a stylized two-period version of the canonical equilibrium default model used to integrate sovereign risk in the analysis of aggregate fluctuations. We then show how the infinite-horizon version of the model (presented in the appendix to this chapter) can account quantitatively for the distinctive features of business cycles in economies with sovereign default risk. In the default model, both the interest rate and the sensitivity of the interest rate to borrowing increase in bad times, generating the countercyclical trade balance and a volatility of consumption higher than the volatility of income that is characteristic of economies facing significant default risk..

24.3 A stylized two-period default model

As Eaton and Gersovitz (1981), we focus on a small-open economy model with strategic defaults. Following Aguiar and Gopinath (2006) and Arellano (2008), an extensive literature uses the infinite-horizon version of this model for the quantitative analysis of aggregate fluctuations. We first present a stylized two-period model that is amenable to an analytical characterization of how default risk constrains the sovereign's choices. We then proceed to presenting the infinite-horizon version of the model used for quantitative analysis.

The small-open economy is populated by a continuum of identical households (who do not make any choices) and a benevolent government that maximizes households' welfare. That is, the domestic economy is modeled as a single agent that we call the government. The economy lasts for two periods, $t = 1, 2$. The economy is endowed with Y_2 goods in period 2.[6] The endowment realization Y_2 is stochastic. Let F and f denote the c.d.f. and density functions of Y_2, with $f(Y_2) > 0$ for all $Y_2 \geq 0$. We assume the economy is endowed with $Y_1 = 0$ goods in period 1, and the government enters period 1 with zero debt. This ensures the government needs to borrow in period 1.

The government's objective is to maximize the social welfare function

$$u(C_1) + \beta \mathbb{E} u(C_2),$$

where u denotes the consumers' utility function, C_t denotes consumption in period t, $\beta \in (0, 1)$ denotes households' discount factor, and $\mathbb{E}$ denotes the expectation operator. The utility function satisfies $u' > 0$, $u'' < 0$, and the Inada condition.

In period 1, the government can borrow to finance consumption. A bond issued in period 1 promises to pay one unit of the single good in period 2. Bonds are issued to foreign, risk-neutral investors that operate in a competitive market. To simplify notation, we assume these investors discount future payments at a rate of 1, i.e., they ask for a zero interest rate compensation when purchasing bonds.

The government cannot commit to repaying its debt in period 2. If the government defaults, it does not pay its debt but loses a fraction $\phi > 0$ of Y_2. For sovereign debt to materialize in equilibrium, it must be more costly for a sovereign to default than to pay back its debt for at least some states of the economy. Without a cost of defaulting, the government would always default and, anticipating this, investors would never lend. At the same time, for sovereign defaults to occur in equilibrium, there must be some states for which it is more costly for the government to pay back its debt than to default. Most of the quantitative default literature presents a simple formulation of the cost of defaulting calibrated to match properties of the sovereign spread in the data. Measuring the effect of default on income is difficult because of a reverse causality problem: while defaults could have a negative effect on income, low income (and the expectation of low income in the future) may push the government to default. In contrast, sovereign spreads are easily measured, and the dynamics of sovereign spreads in the simulations of the quantitative model are determined by the assumptions on the default cost.[7]

[6] Each household is endowed with $y_2 = Y_2$ goods.

[7] The literature discusses as possible costs of sovereign defaults financial penalties in the form of higher borrowing costs and/or exclusion from debt markets for defaulting countries (Cruces and Trebesch, 2013),

As is standard in the literature, we use the Markov-Perfect equilibrium concept. This is, we assume that in period 1, the government cannot commit to its period-2 default decision. Thus, one may interpret this environment as a game in which the government making decisions in period 1 is a player who takes as given the default strategy of its future self, who will decide in period 2.

Equilibrium borrowing and default

To characterize the equilibrium in this model, we solve it backwards.

Period 2

Let B_2 denote the number of bonds issued by the government in period 1 and that mature in period 2. Let D_2 denote the default decision in period 2, where $D_2 = 1$ (0) indicates that the government defaults (repays). Since there is no borrowing in period 2, consumption in period 2 satisfies

$$C_2 = \begin{cases} Y_2 - B_2 & \text{if } D_2 = 0, \text{ and} \\ Y_2 - \phi Y_2 & \text{if } D_2 = 1. \end{cases} \tag{24.3}$$

We assume that the government repays when it is indifferent between repaying and default. The government decides whether to default in period 2 maximizing utility $u(C_2)$. Therefore, the government picks the highest possible value for C_2, implying that it defaults if and only if the cost of defaulting (ϕY_2) is below than the cost of paying its debt (B_2). Let $\hat{D}_2$ and $\hat{C}_2$ denote the equilibrium default and consumption decisions in period 2. It follows that

$$\hat{D}_2(B_2, Y_2) = \begin{cases} 1 & \text{if } Y_2 < B_2/\phi, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{24.4}$$

and

$$\hat{C}_2(B_2, Y_2) = Y_2 - \hat{D}_2(B_2, Y_2)\phi Y_2 - [1 - \hat{D}_2(B_2, Y_2)]B_2.$$

Period 1

In period 1, investors bid to purchase bonds. Since we assumed investors do not discount the future, they bid a price q_1 that coincides with what they expect to recover from each bond. Since each sovereign bond promises to pay one unit of the good in period 2, what lenders expect to recover is this one unit multiplied by the probability of repayment. Formally, the equilibrium bond price satisfies

$$q_1(B_2) = \mathbb{E}\left[1 - \hat{D}_2(B_2)\right] = Pr(Y_2 < B_2/\phi) = 1 - F\left(B_2/\phi\right). \tag{24.5}$$

direct sanctions and trade costs (Asonuma, Chamon, and Sasahara, 2016), reputational spillovers (Cole and Kehoe, 1998 and Amador and Phelan, 2021), and domestic financial and political costs (Broner, Martin, and Ventura, 2010). As presented in the Appendix, the quantitative default model often assumes that the defaulting country is excluded from debt markets for a stochastic number of periods. This assumption does not have a significant effect on the properties of the model . Bulow and Rogoff (1989) show that assuming financial autarchy as the only cost of defaulting would not be enough to support debt in equilibrium.

In a rational expectations equilibrium, the repayment probability expected by lenders is determined by the government's equilibrium default decision, specified in equation (24.4). The price q_1 is an equilibrium because: i) if investors offered to buy bonds at a higher price than $q_1(B_2)$, they will earn negative profits in expectation, and ii) if they offered a lower price than q_1, they will be outbid by other investors.

Equation (24.5) illustrates how, in equilibrium default models, borrowing choices available to the government are restricted by the limited commitment problem. The bond price decreases with the number of bonds issued. The government's inability to commit to repaying in period 2 limits how much it can borrow in period 1.[8]

At the beginning of period 1, the government chooses how much to borrow taking as given its own equilibrium default decision in period 2 (equation 24.4) and the implied bond price schedule (equation 24.5). Thus, the government's problem in period 1 consists of

$$\max_{B_2 \geq 0} \left\{ u\left(C_1\right) + \beta \mathbb{E}\left[u\left(\hat{C}_2(B_2, Y_2) \right) \right] \right\}$$

subject to

$$C_1 = B_2 q_1(B_2).$$

Equilibrium borrowing $\hat{B}_2$ satisfies

$$u'\left(C_1(B_2)\right) \left[q_1(B_2) + B_2 \frac{dq_1(B_2)}{dB_2} \right] = \beta \mathbb{E}\left[u'\left(\hat{C}_2(B_2, Y_2) \right) \left[1 - \hat{D}_2(B_2, Y_2) \right] \right]. \tag{24.6}$$

Equation (24.6) shows how the Euler equation in a model with defaults is different from the one in a model without default. The current marginal benefit of issuing an extra bond depends on the value of increasing consumption in period 1 ($u'(C_1)$), and the marginal revenue from issuing an extra bond ($q_1 + B_2 dq_1/dB_2$). In a default model, this marginal revenue depends on the derivative of the bond price because when the government issues an extra bond, it increases the default probability on all the bonds issued in period 1, not only the default probability on the extra bond issued. The right-hand side represents the expected repayment cost of issuing an extra bond. This cost is determined by the additional resources the government would need to take from households in period 2 to repay B_2, reducing consumption in repayment states (when $\hat{D}_2 = 0$). Since we assumed the government writes off the entire debt stock when it defaults, consumption in default states does not depend on B_2.

24.4 Simulations using a quantitative default model

This section presents simulation results from a quantitative infinite-horizon version of the default model (in which the government faces income shocks and can choose whether to default in every period) calibrated using data from Mexico as reference. We first describe

[8] Note that following the majority of the literature and to be consistent with the most commonly used sovereign debt instruments, we are assuming the government issues non-contingent debt. In our stylized model, allowing the government to issue income-contingent debt would expand its borrowing opportunities and eliminate defaults. Roch and Roldán (2023) discuss limitations of state-contingent debt instruments.

the fully dynamic environment here and define the equilibrium we will study (using dynamic programming). This model is a straightforward extension of the two-period model above, but solving it requires numerical methods and we explain different methods in Appendix 24.A, along with a description of the details of the calibration.

24.4.1 The environment

Preferences. The sovereign's preferences are given by

$$\mathbb{E}_t \sum_{j=t}^{\infty} \beta^{j-t} u\left(C_j\right).$$

The utility function is strictly increasing and concave.

Technology. We assume that the economy receives a stochastic endowment stream of a single tradable and perishable good. The economy's endowment is denoted by $Y \in \mathbb{R}_{++}$. The endowment process follows:

$$\log(Y_t) = (1-\rho)\,\mu + \rho \log(Y_{t-1}) + \varepsilon_t,$$

with $|\rho| < 1$ and $\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$.

Debt Structure. The sovereign issues long-term bonds. Since the relevant state space for the sovereign consists of the future stream of debt payments, long-term bonds need to be modeled in a way that maintains tractability. We assume that the sovereign issues perpetuities that promise a deterministic stream of coupons that decreases at an exogenous constant rate δ (Hatchondo and Martinez, 2009; Arellano and Ramanarayanan, 2012). In particular, a bond issued in period t promises to pay $\delta(1-\delta)^{j-1}$ units of the tradable good in period $t+j$, for all $j \geq 1$. Hence, debt dynamics can be represented by the following law of motion:

$$B_{t+1} = (1-\delta)B_t + I_t, \tag{24.7}$$

where B_t is the number of bonds due at the beginning of period t, and I_t is the amount of bonds issued in period t. The advantage of this formulation is that the current debt stock (B_t) is sufficient for predicting the future stream of debt payments (absent future debt issuance). Equation (24.7) is akin to the equation for capital accumulation in the neoclassical growth model. In that model, agents do not need to know when each unit of capital was incorporated into the economy because all capital units depreciate at a constant rate. Equation (24.7) shows that agents in this model do not need to know when each bond was issued because all bond payments decline at the same rate.

Budget constraint. Let q_t denote the price at which the sovereign issues bonds. When the sovereign has access to debt markets, it faces the following budget constraint:

$$C_t + \delta B_t = Y_t + I_t q_t. \tag{24.8}$$

The sovereign finances consumption (C_t) and coupon payments (δB_t) with income (Y_t) and the resources collected from debt issuance ($I_t q_t$).

Defaults. The sovereign cannot commit to repay its debt obligations and can declare a default in any period. When the sovereign defaults, it does so on all current and future debt obligations.[9] The canonical model assumes that the recovery rate for debt in default (i.e., the fraction of the loan that lenders recover after a default) is zero (this assumptions has been relaxed in several studies and is inconsequential for the issues we underscore in this chapter).

The cost of defaulting is that the sovereign is excluded from debt markets for a stochastic number of periods and suffers an income loss $\phi(Y_t)$ in each period in which the economy is excluded from debt markets. As in Arellano (2008) and Chatterjee and Eyigungor (2012), we assume that it is proportionally more costly to default in good times ($\phi(Y)/Y$ increases in Y). They show that this property is important in accounting for spread dynamics.[10]

Bond market. Bonds are priced in a competitive market inhabited by a large number of foreign investors. Thus, the bond price q_t is pinned down by a zero-expected profit condition. Investors are risk-neutral and discount future payoffs at the rate r, representing the global risk-free interest rate and thus the investors' opportunity cost of lending.

Timing. The timing of events within each period is as follows. Firstly, the sovereign and investors observe the income realization. Secondly, the sovereign chooses whether to default on its debt. Thirdly, if the sovereign does not default, it pays current debt obligations and rebalances its debt portfolio.

Equilibrium concept. We use the Markov-Perfect equilibrium concept. The sovereign cannot commit to future default and borrowing decisions. The strategy for the sovereign acting in period t is assumed to depend only on payoff relevant variables, which in this environment consist of debt and income in period t (B_t, Y_t).

24.4.2 Recursive formulation

Let D denote the current-period default decision, where $D = 1$ if the sovereign defaults and $D = 0$ if the sovereign repays. Let V denote the sovereign's value function at the beginning of a period, that is, before the default decision is made. Let V^R denote the value function of a sovereign that has repaid its debt and V^D denote the value function of a sovereign in default. The function V satisfies the following functional equation:

$$V(B, Y) = \max_{D \in \{0,1\}} \{DV^D(Y) + (1 - D)V^R(B, Y)\}. \tag{24.9}$$

If the sovereign defaults, the economy consumes its available resources $y - \phi(Y)$. With probability ψ, the sovereign may regain access to bond markets in each of the following

[9] This is a standard assumption in the literature and is consistent with the observed behavior of defaulting sovereigns and the widespread presence of the acceleration and cross-default clauses.

[10] Mendoza and Yue (2012) shows that this property of the cost of defaulting arises endogenously in a setup in which defaults affect the ability of local firms to acquire a foreign intermediate input good.

periods. Let F denote the conditional cumulative distribution function of the next-period endowment y'. The function V^D satisfies the following functional equation:

$$V^D(Y) = u\left(Y - \phi\left(Y\right)\right) + \beta \int \left[\psi V(0, Y') + (1-\psi)\, V^D(Y')\right] F\left(dY' \mid Y\right). \tag{24.10}$$

Note that the sovereign regains access to debt markets with zero debt and thus, the continuation value of that contingency is $V(0, Y')$.

If the sovereign repays, current consumption is determined by equation (24.8). We avoid a potential discontinuity at $B' = 0$ with the constraint $B' \geq 0$. We verify that this constraint is not binding in the simulations.[11] For any bond price function q, V^R satisfies the following functional equation:

$$V^R(B, Y) = \max_{B' \geq 0} \left\{ u\left(C\right) + \beta \int V(B', Y') F\left(dY' \mid Y\right) \right\}, \tag{24.11}$$

subject to

$$C = Y - \delta B + q(B', Y)\left[B' - (1-\delta)B\right].$$

For investors to break even, the bond price must equal the expected discounted value of debt payments. Namely:

$$q(B', Y) = \frac{1}{1+r} \int \left[1 - \hat{D}\left(B', Y'\right)\right] \left[\delta + (1-\delta) q(\hat{B}(B', Y'), Y')\right] F\left(dY' \mid Y\right), \tag{24.12}$$

where $\hat{D}$ and $\hat{B}$ denote the default and borrowing rules lenders expect the sovereign to follow in the next period. The default rule $\hat{D}$ is equal to 1 if the sovereign defaults and is equal to 0 otherwise. The function $\hat{B}$ determines the debt stock chosen by the current sovereign. The first term on the right-hand side of equation (24.12) equals the expected value of the next-period coupon payment promised in a bond. The second term in the right-hand side of equation (24.12) equals the expected value of all other future coupon payments, which is summarized by the expected price at which the bond could be sold next period.

24.4.3 Equilibrium definition

A Markov Perfect Equilibrium is characterized by

1. Value functions: V, V^R, and V^D,
2. a default rule $\hat{D}$ and a borrowing rule $\hat{B}$, and
3. a bond price function q,

[11] In the simulations, the sovereign typically does not even buy back debt (i.e., it never chooses $B' < (1-\delta)b$). In an environment without income shocks, Aguiar, Amador, Hopenhayn, and Werning (2019) show it is never optimal for the sovereign to buy back debt.

such that:

(a) given $\hat{D}$ and $\hat{B}$, V, V^R, and V^D satisfy functional equations (24.9), (24.10), and (24.11), when the sovereign can trade bonds at q;

(b) given $\hat{D}$ and $\hat{B}$, the bond price function q is given by equation (24.12); and

(c) the default rule $\hat{D}$ and borrowing rule $\hat{B}$ solve the dynamic programming problem defined by equations (24.9) and (24.11) when the sovereign can trade bonds at q.

24.4.4 Results

Table 24.3 illustrates first how the simulations can match the targeted average levels of debt and spreads in the data.[12] Table 24.3 also shows that the default model can account for distinctive features of business cycles in emerging economies: consumption is more volatile than income and the trade balance and sovereign spread are countercyclical.[13]

Figures 24.6 and 24.7 illustrate why the predictions of the quantitative model are consistent with the data for emerging economies. In Figure 24.6, as in the two-period model, the bond price decreases with debt. The bond price also increases with income, implying that the sovereign's borrowing set shrinks when income drops, precisely when the sovereign needs borrowing the most. This occurs because when income takes low values, investors anticipate lowe income and thus stronger default incentives in future periods.[14] The solid dots in Figure 24.6 illustrate how, in equilibrium, the sovereign chooses to issue fewer bonds and at a lower price when the income realization is low (one standard deviation below the mean in the graph). Figure 24.7 shows that when income is low, because of the lower borrowing, consumption drops more than income, accounting for the excess consumption volatility in emerging economies.

Normative implications in economies with default risk. We have shown that the canonical default model offers an empirically plausible micro-founded account of business cycle dynamics in emerging economies. Could this model also be useful for policy design? Note that the implication of the model is that a benevolent government (maximizing the

[12] As discussed in the Appendix, there are two important assumptions for achieving this: (i) the government issues long-term debt (Arellano and Ramanarayanan, 2012; Chatterjee and Eyigungor, 2012; Hatchondo and Martinez, 2009), and (ii) the cost of defaulting increases more than proportionally with income (Arellano, 2008; Chatterjee and Eyigungor, 2012).

[13] Note that Mexico has experienced a moderation in aggregate fluctuations in recent years and thus consumption is as volatile as income in Table 24.3 but it was 24% more volatile than income in the earlier sample used by Aguiar and Gopinath, 2006.

[14] In terms of the two-period model, consider a version of this model with $Y_1 > 0$ and a c.d.f. function for period-2 income that is decreasing with respect to period-1 income, $F(\cdot \mid Y_1)$. For any chosen B_2, a higher income in the first period shifts the distribution of future income (Y_2) to the right, thereby reducing the default probability. As a consequence, the equilibrium bond price

$$q_1(B_2, Y_1) = 1 - F(B_2/\phi \mid Y_1)$$

is an increasing function of period-1 income.

Table 24.3: Business Cycle Statistics: Model and Data

Targeted moments		
	Model	Data
Mean Debt-to-GDP	43	43
Mean $Spread$	3.2	3.2
Non-Targeted moments		
$\sigma(C)/\sigma(Y)$	1.3	1.0
$\sigma(TB/Y)$	0.8	1.3
$\sigma\left(Spread\right)$	1.6	2.2
$\rho(TB/Y, Y)$	-0.7	-0.6
$\rho\left(C, Y\right)$	0.98	0.91
$\rho\left(Spread, Y\right)$	-0.8	-0.4
$\rho(Spread, TB/Y)$	0.9	0.6

Note: The standard deviation of x is denoted by $\sigma\left(x\right)$. The coefficient of correlation between x and z is denoted by $\rho\left(x, z\right)$. Moments are computed using detrended series. Trends are computed using the Hodrick-Prescott filter with a smoothing parameter of $1,600$. Moments for the simulations correspond to the mean value of each moment in 500 simulation samples. We take the last 120 periods (30 years) for each sample without a default episode. Simulation samples start at least five years after a default. Default episodes are excluded to improve comparability with the data. Consumption and income are expressed in logs.

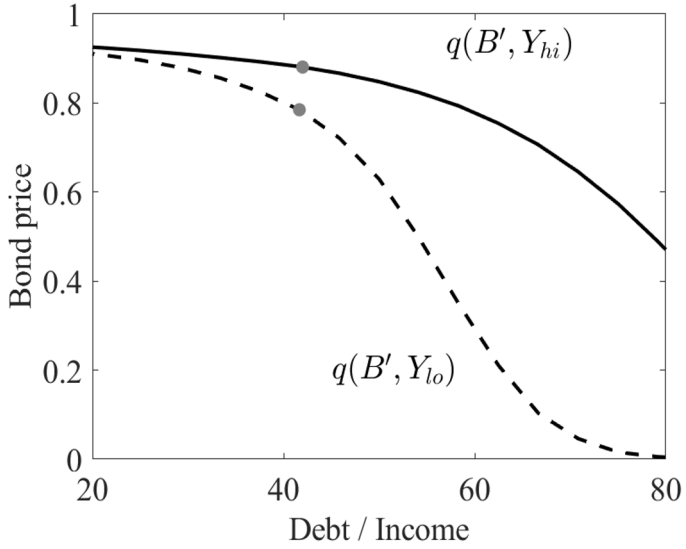


Figure 24.6: Bond price schedule for a low income ($Y_{lo} = E(Y) - \sigma(Y)$) and a high income level ($Y_{hi} = E(Y) + \sigma(Y)$).

Notes: The horizontal axis displays the ratio of debt to average annual income $B'/4E(Y)$ the government could choose. Solid dots correspond to the optimal choices when the government enters the period with the mean debt level in the simulations.

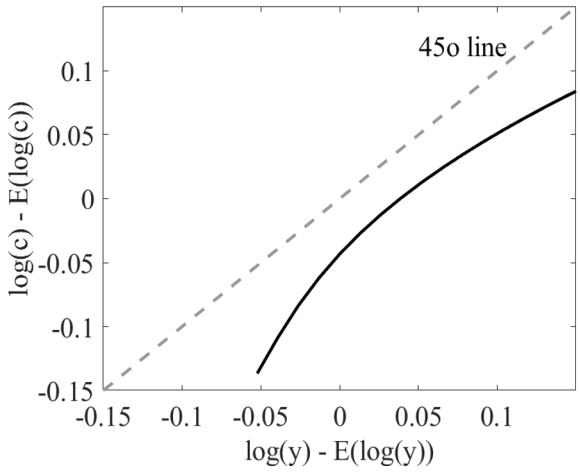


Figure 24.7: Consumption policy under repayment when the sovereign enters the period with the mean debt level in the simulations.

utility of the representative consumer) chooses to face high sovereign risk. At the same time, as discussed in the introduction, having high sovereign risk seems to have significant adverse effects no only on business cycle dynamics but also on the ability of governments to react to large shocks (e.g., COVID-19) and on investment and growth. Given these costs of sovereign risk, why would benevolent governments still choose debt levels that command high default risk?

The answer to this question may be related to time inconsistency problems that for example, also lead governments to choose high inflation. The default model features two time inconsistency problems. First, the government may want to commit to default less in the future (in comparison to how often it chooses to default in equilibrium). The government acting in period t could benefit from restricting defaults in period $t+1$ (and with long-term debt, in every future period) because this would expand the borrowing set in t (i.e., would increase bond prices in t). An improved borrowing set would allow the government to better exploit the gains from trade derived from its relative impatience as the calibration of quantitative default models typically assumes that the government discounts the future more than lenders. With β being the government's discount factor and r be the interest rate representing the lenders' opportunity cost of lending, the standard calibration assumes that $\beta(1+r) < 1$. Bolton and Jeanne (2009) study a model in which issuing debt that is more costly to restructure can be optimal, as this would deter future defaults. As mentioned before, governments in emerging countries do issue debt in foreign financial centers and include covenants in debt contracts to make their debt more difficult to restructure. Mateos-Planas et al. (2025) find that for standard calibrations of the default model, if the government could commit to its future default policy, it would commit to not defaulting. They also show that for less standard calibrations it is possible to obtain default under commitment, but default is very rare. This is, the time inconsistency problem due to the government's inability to commit to future default decisions accounts for the majority if not all equilibrium sovereign risk. Hatchondo et al. (2022) show that committing to a no-default rule would generate large welfare gains for the government, but the temptation to break this rule would also be large.

The second time inconsistency problem stems from the government's inability to commit to future borrowing levels. This problem only arises when we assume that the government can issue long-term debt (Arellano and Ramanarayanan, 2012; Chatterjee and Eyigungor, 2012; Hatchondo and Martinez, 2009). The mechanism behind this time consistency problem is similar to the one described above: by committing to lower future borrowing levels, the government could increase current bond prices. For example, by restricting borrowing in $t+1$, the sovereign can reduce default probabilities in $t+2$ (and after) and expand the borrowing set in t. Note that with one-period debt, the default probability in $t+2$ does not affect bond prices in t (which only depend on the default probability in $t+1$) and thus, this time inconsistency problem does not arise (see Aguiar and Amador, 2019 for a formal proof).

Assuming differentiability and that the government issues perpetuities with coupon payments declining at the rate δ, the time-inconsistency problem in borrowing policies is apparent from the period-t optimality condition:

$$u'(C_t)\left[q_t+\frac{\partial q_t}{\partial B_{t+1}}\left[\underbrace{B_{t+1}-(1-\delta)B_t}_{\text{Bonds issued at } t}\right]\right]=\beta\mathbb{E}\left[V_1\left(B_{t+1},Y_{t+1}\right)\mid Y_t\right], \tag{24.13}$$

where V_1 denotes the derivative of the government's value function with respect to debt. Note that the optimality condition in equation (24.13) resembles the one for the two-period model in equation (24.6).

Equation (24.13) illustrates how, when the sovereign borrows in t, it takes into account the negative effect that an extra bond sold has on the price at which it sells bonds, q_t. However, the sovereign deciding in t does not take into account how borrowing in t affects the price of the bonds it issued in periods before t. Of course, a government deciding in period t should not take into account how its borrowing affect the price of bonds it issued in previous periods. However, a government that could commit to an optimal borrowing plan before t, would take into account how borrowing at t would affect the price of bonds (and thus utility) before t (as the government who could commit to future default plans would take into account the effect of future default on past prices).

As discussed in the introduction, a large and increasing number of countries are implementing fiscal rules to limit borrowing by future governments. Hatchondo et al. (2022) use the default model to discuss the design of fiscal rules. Chatterjee and Eyigungor (2015) and Hatchondo, Martinez, and Sosa Padilla (2016) show how alternative bond contracts could help mitigate the time inconsistency problem in borrowing policies.

Chapter 25

Sustainability

John Hassler, Per Krusell, and Conny Olovsson

25.1 Introduction

With climate change surfacing as one of the major challenges of modern society, many economists have of course taken an interest in the subject. This includes macroeconomists. In a very clear sense, fighting global warming involves aggregates. Climate change itself is global in nature—there is warming all around the world—and it is fundamentally caused by aggregate CO_2 emissions, which are a natural byproduct of the use of fossil fuel as an energy source. Since fossil fuel, since long, has constituted our main source of energy, this is an aggregate issue, at least to the extent energy use is important. The services provided by energy generate a cost that roughly amounts to 5% of GDP globally. This is not a small amount and, moreover, there are many reasons to believe that sharp reductions in energy use might cause significant recessions; the oil-price hikes during the 1970s are often argued to have had such an effect. Relatedly, suppose our world leaders decide to enact policies that are meant to contain global warming. What, then, would be good such policies? Climate change is a long-run issue and direct evidence on the pros and cons of different policy choices in this arena, not only in the short run but also in the long run, is not available. For that reason, macroeconomic modeling appears useful. Thus, one of the main purposes of the present chapter is to develop some simple tools with the goal of assessing the impact of different policy paths on the climate and on our welfare. Fortunately, rather minor departures from our basic macroeconomic models are required for this endeavor, so we will be able to build significantly on earlier chapters.

A broader, global issue is that of *sustainability*: do humans, those presently alive and those of the future, have enough of what they need, for survival and for living happy lives? To what extent should we economize on natural resources of various kinds? Which, if any, broad policy tools should be considered? To be more concrete will we need to pursue policies that aim to limit world output growth or even decrease output: *de-growth*? These are questions that are obviously hard to answer definitively, but the chapter will at least begin to address them with available tools, drawing mostly on those taught in this textbook and on classic results in public/environmental economics.

The chapter begins in Section 25.2 with a general, but short, discussion of sustainability,

placing "the economy"—the subject of the field of economics—in the broader context of natural resources and how the field of economics has, or has not, made contact with natural resources. This connection will receive attendion in the second part of the chapter—in Section 25.5—where we discuss basic theory as well as technological change in this context. Before that, however, in Section 25.3, in order to build toward a full analysis of climate change, we explain "how the climate works": we offer a summary of the natural-science knowledge required for understanding how to analyze global warming from an economics perspective. This section also summarizes some basic facts about fossil fuel supplies as well as the rapidly growing literature on the economic—broadly defined—consequences of climate change. With this background, we turn to our main analysis of the climate-economic interactions in Section 25.4. In this chapter, we employ a summary of how the climate works in the form of a system of dynamic equations and add these equations to our macroeconomic model and, hence, obtain an *Integrated Assessment Model* (IAM). The word "integrated" here thus refers to the combination of the climate sciences and the economic sciences; "assessment" refers to using the model as a laboratory: to evaluate how different policy actions work their way toward climate outcomes and economic outcomes. We also simulate the model to illustrate its usefulness for policy evaluation.

25.2 The economy and the environment

The term "environment" is a standard term in microeconomic theory to represent the basic elements of an economy (preferences, technology, etc.). Here, we use it to refer to "nature": what planet Earth (and the solar system) gives us access to for free. Today, there is clearly a movement—especially in advanced economies—to preserve the environment in a variety of ways, and this chapter will make contact with some of the issues pertinent to this movement.

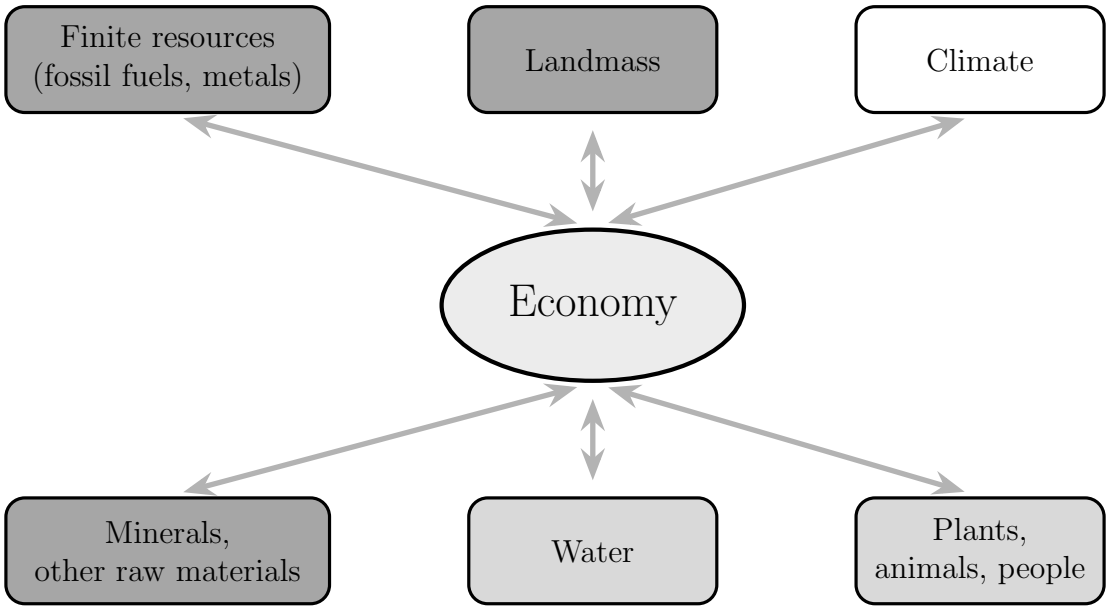


Figure 25.1: The economy and the environment. Darker shades of grey reflect a higher degree of property rights.

Figure 25.1 illustrates and depicts a moment in time, say a year. The "economy" consists of households, firms, governments, and other organizations, along with their preferences and production possibility sets (as well as information), as well as capital stocks and infrastructure built up in the past. The economy in turn interacts with Earth's environment, which we separate schematically into different components. On the top left we have land mass, which of course is used for a number of economic activities. Land mass has the darkest shade: it is the category which we guess has the highest degree of property rights attached to it. Second, we have finite resources, such as fossils (which can be used as fuels), minerals, and so on. We approximate these to be non-renewable, even though fossils and minerals are reproduced but at a time scale that is so long that we can consider these resources to be in fixed supply. So we use up these resources over time but can also recycle. Much of these resources are owned by someone. Ownership is relevant since Coase's insights can potentially be used: the owner could themselves manage the resources over time in a purposeful way as a function of market prices, and so it is not obvious at the outset that these resources are overconsumed. The management of natural resources in finite supply is discussed in Section 25.5 below.

Our "global climate" is the average weather (temperature, wind speeds, precipitation, etc.) across the globe and it is determined by a number of factors exogenous to humans, but also—as we will discuss in detail in Section 25.3—by the amount of greenhouse gases in the atmosphere and some of these are directly caused by human economic activity. It is not possible to attach property rights to the atmospheric greenhouse gas concentration mainly because this concentration is even around the world: any emissions spread around the Earth very fast. Thus the global climate is in the lightest in the picture. Of course local climate differences are tied to local land ownership and so local climate can be "transacted", for example through tourism.

Human activity also crucially uses water—oceans, lakes, and rivers—in a variety of ways, and we can affect the amount of water available, too. Then there are forests, plants, etc., i.e., living organisms; like water, these are inputs into production and consumption and relevant for leisure activities. The concept *biodiversity* refers to the amount of variety within this category.

Finally, there is solar energy: it is a form of radiation and gives us energy without which we would not survive. The solar energy we have access to is a flow. It does vary somewhat over time and is an exogenous determinant of our climate.[1] Solar energy can be captured to some extent by land owners, hence giving it a relatively dark color.

The amount and quality of water, biodiversity, and land are all endogenous to human activity as well as to the climate. In Section 25.3.5 below, we discuss how climate change causes "damages" by reducing the value (to humans) of elements within these categories.

The chapter will first discuss the third category, the climate, and its interaction with the economy. To study this interaction, we employ so-called Integrated Assessment Models—IAM from now on. Even though the details may differ substantially between different IAMs, they all consist of three main building blocks: an economy, a carbon cycle, and a climate module. The term "integrated" refers to the fact that the building blocks affect and interact

[1]The solar energy flow itself cannot be affected by humans but the part of it that reaches our planet can, at least in principle. One form of *geoengineering* involves sending up giant "parasols" into space with the purpose of reflecting sunlight before it reaches us.

with each other, as illustrated in Figure 25.2.

Starting with the economy, this block typically consists of a growth model with households and firms. The firms use fossil energy to produce, thus drawing down on an element of the first environmental category in the previous picture. This gives rise to CO_2 emissions as a by-product when the fossil fuel is burnt. The generated carbon dioxide then enters the next block, i.e., the carbon circulation (or carbon cycle).

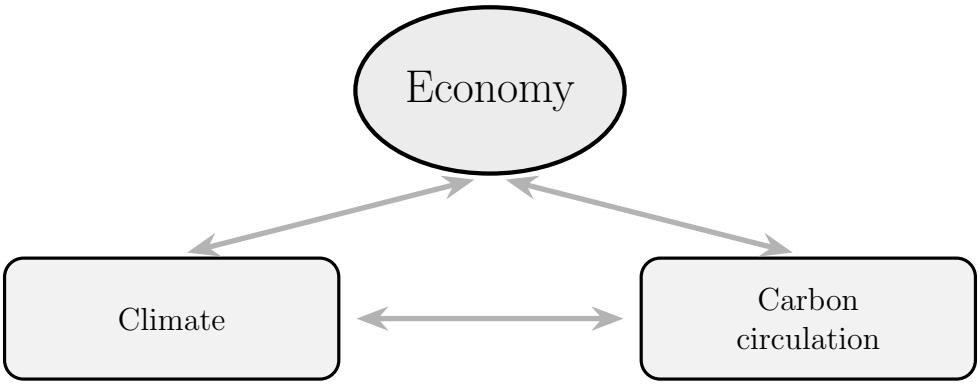


Figure 25.2: Integrated Assessment Models, IAMs, consist of three building blocks: an economy, a representation of the carbon cycle, and a climate module.

As we will discuss below, carbon (dioxide) circulates in a continuous flow between ground, air and sea, but what matters for climate change is the amount of CO_2 in the atmosphere. We will discuss this phenomenon too below. A time path for the stock of CO_2 in the atmosphere is thus the output from the carbon cycle block that enters the climate block where it affects the balance between the incoming energy from the sun and the outgoing heat radiation from Earth to space. Finally, the climate affects the economy through climate-related damages. It is this clockwise rotation that typically is modelled in IAMs, but as is indicated in Figure 25.2, in general, the arrows go both ways.[2]

The IAM we will build below is a neoclassical growth model—as climate change plays out on the time scale of decades and centuries, a long-run perspective is necessary—together with a climate module and a carbon cycle module, each in the form of a set of dynamic equations. To understand the nature of the equations we add to the growth model, we will go through the natural-science basics next, in Section 25.3. This is a textbook for economists, however, and it is possible to skip over this section, thus taking as given the extra equations we later use. Sections 25.3.4 and 25.3.5 may still be relevant: they discuss the fossil reserves left on the planet, along with the costs of extracting them, and how we measure damages (and how large they are estimated to be).

[2]For example, the climate affects the carbon circulation because warmer oceans absorb carbon less easily; increased CO_2 concentration in the oceans makes them more acid, which has negative side effects on our economy, broadly defined; and the economy's burning of fossil fuel directly affect the climate by emitting soot particles into the air, which affects warming.

25.3 Climate change: natural-science background

Let us start with a basic primer on the interaction between the economy and the climate. We will then go into detail on each topic in the subsection below. For over 100 years, it has been known that the amount of carbon dioxide (CO_2) in the atmosphere affects the Earth's *energy balance*. This balance captures the difference between the inflow of energy—in the form of solar radiation—and the outflow of energy, with 1/3 of the outflow consisting of direct reflection and 2/3 consisting of heat radiation.

The carbon dioxide does not affect the energy inflow, because solar radiation passes right through the CO_2. The outflow of energy, however, is affected by the CO_2 in that it makes it more difficult for heat to radiate back into outer space. If the system is initially in balance, an increase in the amount of CO_2 in the atmosphere will lead to a surplus in the energy balance, and as a result, the temperature will rise.

This phenomenon is known as the greenhouse effect. The name is motivated by the fact that it acts exactly like a greenhouse, where solar radiation passes right through the glass of the house whereas the heat gets trapped inside the greenhouse. The greenhouse effect is not entirely a bad thing. Without greenhouse gases in the atmosphere, the Earth would, in fact, be covered in ice and the average surface temperature would be around minus 20°C instead of the close to 15°C that currently prevails.

The emission of greenhouse gases thus leads to climate change and global warming and these changes then affect our economies. Even though these effects from climate change—referred to as *damages*—will vary across space and take many different forms, the aggregate result is projected to be negative for human welfare. We will discuss climate-related damages in detail below, but for now it suffices to state that the damages will be the result of more intensive storms, increased frequency of forest fires, rising sea levels and floods, and so on. These changes in the climate are expected to destroy property and capital, kill and injure people, and may lead to mass migration, political instability, and economic crises.

Because it is generally free to emit greenhouse gases into the atmosphere, even though these emissions gives rise to economic costs, the emissions constitute negative externalities. When market failures such as externalities are present, policy is needed to restore efficiency.

We now describe the basics of the climate system (Section 25.3) and the carbon cycle (Section 25.3.2). In Section 25.3.3 we then describe a relatively recent finding that characterizes the combined carbon-cycle and climate systems into an approximately constant *carbon-climate response* (CCR) to accumulated emissions. Further, in Section 25.3.4, we discuss the world supplies of different kinds of fossil fuel and, finally, in Section 25.3.5 we briefly cover the literature on damage measurements.

25.3.1 The climate system

Let us now describe how CO_2 affects the dynamics of the average atmospheric surface temperature, denoted T and measured as deviations from the pre-industrial temperature. The energy balance is also defined with respect to the pre-industrial situation. The greenhouse effect has a positive effect on the energy balance by reducing the energy outflow. Such changes to the energy balance are called *forcing*.

A second term in the energy balance is due to the fact that hotter objects radiate more

energy. Thus, as Earth gets warmer, more energy is radiated to space. We approximate this effect to be proportional to T, with a coefficient κ. Finally we approximate the change dT/dt as being proportional to the surplus in the energy balance with a coefficient σ_1. For a given forcing f time path, the evolution T (in continuous time) is

$$\frac{dT(t)}{dt} = \sigma \left(f(t) - \kappa T(t) \right). \tag{25.1}$$

The coefficient κ incorporates *feedback effects*, such as the phenomenon that melting ice reduces the direct reflection of incoming sunlight and the effects of increased temperature on cloud formation. The sum of all the feedbacks is highly uncertain and hence the value of κ is uncertain. After an increase in forcing to a constant level f, the temperature increases and the system eventually reaches a new steady state when $T(t) = f/\kappa$.

Human activity has substantially increased the concentration of greenhouse gases in the atmosphere. The dominant greenhouse gas that stems from human activity is CO_2. It is currently responsible for about two thirds of the forcing from greenhouse gases in the atmosphere.[3] Since the beginning of the Industrial Revolution, we have increased the concentration of CO_2 in the atmosphere by around 50 per cent. Humans also emit particles and aerosols that have a direct effect as well as effects on cloud formation. In sum, these effects are likely cooling, but the uncertainty here is large.

So far, we have learned that higher concentration of CO_2 in atmosphere reduces the outgoing flow of energy in the form of infrared radiation. This phenomenon is well approximated by a logarithmic function (Arrhenius greenhouse law, 1896). If we denote the stock of carbon dioxide in the atmosphere by S, and the pre-industrial level (i.e., the level before the Industrial Revolution) by S_0, then forcing due to carbon dioxide is well approximated by

$$f_{CO_2} = \frac{\eta}{\log 2} \log \frac{S}{S_0}. \tag{25.2}$$

An often used value of η is 3.7. Combining the steady state of (25.1) with (25.2), we obtain the steady-state temperature associated with a given amount S:

$$T\left(f(S)\right) = \frac{\eta}{\kappa} \frac{1}{\log 2} \log \left(\frac{S}{S_0} \right). \tag{25.3}$$

The ratio η/κ is labelled the *equilibrium climate sensitivity* (ECS), and it measures how much the long-run temperature will increase from a doubling of the CO_2 level. The IPCC's (Intergovernmental Panel on Climate Change) 6th report states that the ECS is "likely" between $2.5°C$ and $4°C$, with a"best" estimate of 3°C. A 90-% interval is 2°C–5°C. The IPCC has revised their estimate of the ECS several times. The interval reported in the 6th report is more narrow than that in the 5th report where it was given as being between 1.5°C and 4.5°C, with no best estimate given.

Equation (25.1) does not take into account the fact that the heating of the atmosphere involves interaction with the oceans. In particular, the oceans change temperature quite

[3]However, by far the most important cause of the greenhouse effect is water vapor, but water vapor does not derive directly from human activity.

slowly, creating a long-lasting but not permanent cooling effect on the atmosphere. Denoting the ocean temperature by T^L (again measured as a deviation from its preindustriual level), a difference between the atmospheric and ocean temperatures will create a third term in the energy balance for the atmosphere. This term is a the flow of energy from the atmosphere to the ocean and is approximately proportional to the difference between the two temperatures, with coefficient σ_2. Finally, we need to specify the law-of-motion for T^L and we assume it to be proportional to the temperature difference $T - T^L$ (and thus to the energy flow from the atmosphere to the oceans), with coefficient σ_3. Discretizing the resulting system yields

$$T_t = T_{t-1} + \sigma_1 \left(\frac{\eta}{\log 2} \log \left(\frac{S_{t-1}}{S_0} \right) - \kappa T_{t-1} - \sigma_2 \left(T_{t-1} - T^L_{t-1} \right) \right) \tag{25.4}$$

and

$$T^L_t = T^L_{t-1} + \sigma_3 \left(T_{t-1} - T^L_{t-1} \right). \tag{25.5}$$

It is straightforward to show that for a given S, this system has a steady state that is also given by equation (25.3). However, since an empirically reasonable parameterization requires σ_3 to be much smaller than σ_1, convergence to the steady state is slower than for equation (25.1).

25.3.2 The carbon cycle

The strength of the greenhouse effect depends on the amount of CO_2 in the atmosphere. CO_2 is emitted into the atmosphere and then circulates between three main reservoirs—the atmosphere, the surface oceans and biosphere, and the deep oceans—and we refer to this system as the *carbon cycle*. Denoting the stock of CO_2 in the surface oceans and biosphere and in the deep oceans by S^U and S^L, respectively, the carbon cycle block can be formulated as follows.

$$\begin{aligned} S_t - S_{t-1} &= \phi_{12} S_{t-1} + \phi_{21} S^U_{t-1} + E_{t-1} && (25.6) \\ S^U_t - S^U_{t-1} &= \phi_{12} S_{t-1} - (\phi_{21} + \phi_{23}) S^U_{t-1} + \phi_{32} S^L_{t-1} && (25.7) \\ S^L_t - S^L_{t-1} &= \phi_{23} S^U_{t-1} + \phi_{32} S^L_{t-1}, && (25.8) \end{aligned}$$

where E denotes CO_2 emissions.

The above specification is used in Nordhaus's RICE model. Folini, Kubler, Malova, and Scheidegger (2021) show that the carbon-cycle model above closely replicates the mean behavior of the most advanced Earth System Models (CMIP5), if the parameters are chosen appropriately.[4] The parameter values suggested by Folini et al. (2021) are $\phi_{12} = 0.053$, $\phi_{21} = 0.0536$, $\phi_{23} = 0.0042$, and $\phi_{32} = 0.001422$, when a period is one year. The initial stocks are set to $S_{2015} = 850$, $S2015^U = 765$, and $S^L_{2015} = 1799$.

Note the important difference between measuring emissions in units of CO_2 and carbon units. One kg of carbon corresponds to approximately 3.67 kg of CO_2.[5]. The distinction

[4] CMIP5 stands for Coupled Model Intercomparison Project, Phase 5; it is used by the IPCC and is considered to be the state-of-the-art model in climate science.

[5] A mole of carbon atoms weighs 12 grams, whereas a mole of oxygen weighs 16 grams and we obtain $(2 * 16 + 12)/12 \approx 3.67$

sometimes causes confusion since, e.g., proposed policy tax rates are sometimes expressed per ton of CO_2 and sometimes per ton of carbon.

An alternative to a two-dimensional, linear system, is a one-dimensional reduced form for excess carbon, S:

$$S_t - S_0 = \sum_{t=0}^{\infty}(1 - d_s)E_{t-s}, \tag{25.9}$$

where the key feature again is linearity.[6] Further simplifying approximations can be made with reference to the key characteristics of the carbon cycle as described by the IPCC and Archer (2005). Specifically, about 20-–25 per cent of all carbon dioxide emitted stays in the atmosphere for up to over a thousand years. About 50 per cent disappears in a few decades. The rest takes a few centuries to circulate onwards to the oceans (where it contributes to acidification). Thus,

$$1 - d_s = \phi_L + (1 - \phi_L)\phi_0(1 - \phi)^s \tag{25.10}$$

is proposed in Golosov, Hassler, Krusell, and Tsyvinski (2014) to match these features directly, with the share of emissions that remains in the atmosphere forever represented by ϕ_L, the share that leaves the atmosphere within a period given by $1 - \phi_0$, and the remainder $(1 - \phi_L)\phi_0$ depreciating geometrically at rate ϕ. For a decadal time scale, the calibration $\phi_L = 0.2$, $\phi_0 = 0.393$, and $\phi = 0.0228$ matches the data quite well.

25.3.3 Constant carbon-climate response and the carbon budget

A key feature of the system described by equations (25.4)–(25.8) is that T_t is approximately proportional to cumulative emissions from pre-industrial times until t. This implies that temperature keeps increasing as long as emissions continue and stay constant at the level it has then reached for at long time after emissions have come to an end. The latter is due to two opposite effects that neutralize each other in equation (25.4). On the one hand, CO_2 slowly leaves the atmosphere, reducing forcing (the first term within parenthesis). On the other, the oceans slowly get warmer, reducing their cooling effect (the third term). These terms approximately balance on a centennial scale. These features also hold for the advanced climate models used by IPPC. Formally, we have

$$T_t \approx \sigma_{CCR} \sum_{s=0}^{t} E_s,$$

where T denotes the global mean surface temperature, E global emissions and σ_{CCR} the constant of proportionality that is referred to as the carbon-climate response (CCR), or the Transient Climate Response to Cumulative Carbon Emissions (TCRE). Note of course that this finding is one that combines the climate and carbon modules and generates a direct mapping from emissions to climate.

In the latest (6th) IPCC report, σ_{CCR} is stated as "likely" being between 1.0°C and 2.3°C per 1,000 GtC, which corresponds to 0.27°C–0.63°C/TtCO$_2$).[7] The relationship between

[6] The word "excess" here refers to the level above that which prevailed before the Industrial Revolution.

[7] The term likely is explained to be interpreted as a 2/3 confidence interval.

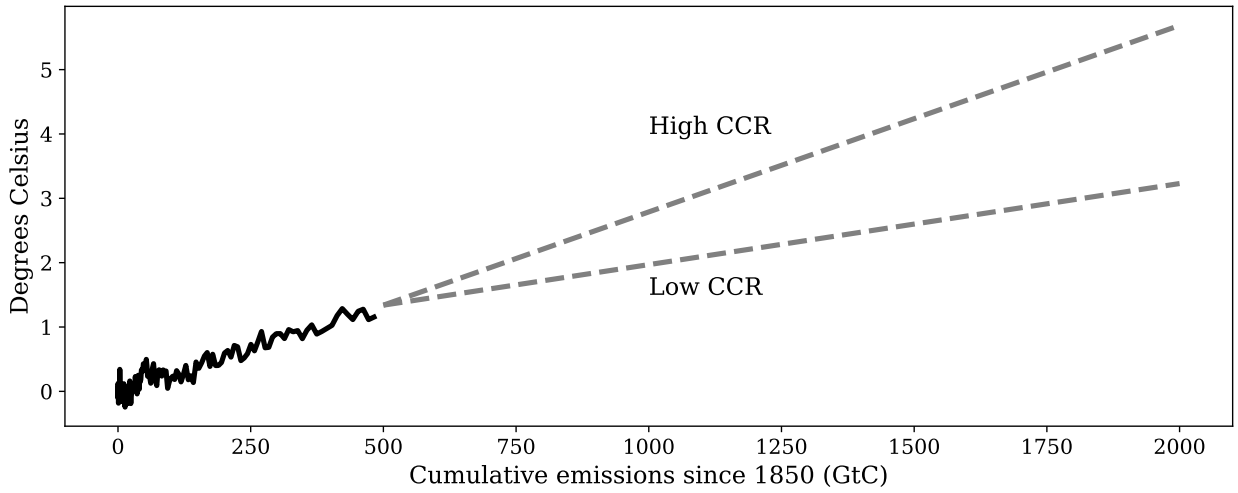


Figure 25.3: Cumulative carbon emissions and the global mean temperature. Solid line: historical data 1850-2022. Dashed lines: forecasts based on constant CCR.

cumulative emissions and the change in global mean temperature based on these results is represented in Figure 25.3. The solid line represents data, i.e., observed cumulative emissions and temperatures from 1850 to 2022. As we see, apart from substantial short-run natural variation, the historical relation is approximately proportional. The dashed lines represent forecasts of the future relationship based on the two endpoints of the uncertainty interval provided by IPCC.[8]

Several valuable insight can be obtained from the finding that the CCR is approximately constant. One is irreversibility: the relation gives temperature at point in time as a function of total past emissions. Hence, there is no depreciation, despite the fact that the carbon cycle fundamentally builds on depreciation, with carbon emitted to the atmosphere, where it acts as a greenhouse gas and causes warming, slowly moving into other sinks where it does not cause warming directly. Thus the conclusion is that if warming is to be stopped, emissions need to stop. Second, a constant CCR also suggests a notion of a *carbon budget*: given that we wish to maintain warming, by a given date, by a certain target amount, we can use the σ_{CCR} to compute how much can be emitted before that date in order to meet the target. Of course, the uncertainty about the CCR coefficient makes this computation somewhat problematic. Up until 2022, we had burnt around 650 gigaton (GtC) of carbon. If the CCR is 1, the warming caused by all historic CO2 emissions is 0.65 x 1 = 0.65°C. To cause a warming of 1.5°C, another 850 GtC would need to be emitted, which would take another 85 years with the current emissions rate. If, on the hand, the CCR is at the upper end of its likely range at 2.3, the 650 GtC already emitted cause a warming of 2.3 x 0.65 = 1.5°C.[9] Third, the constant CCR means that if a target is set for a certain future date, and a carbon budget has been correspondingly calculated, then there are many paths that could

[8]To account for the effect of other greenhouse gases, we have, somewhat arbitrarily, assumed the same proportional relation between this and the effect of CO_2 as has been previously observed. This amounts to multiplying σ_{CCR} by 1.26.

[9]Note that this calculation does not take into account the effects of other greenhouse gases. These effects, however, are much less long-lasting than those of CO2-emissions.

attain the target, ranging from high emissions early and gradual reductions over time to the reverse kind of time path, with radical reductions early and less and less radical reductions over time—all subject to the same total accumulated emissions. These alternative ways of reaching the target would, however, not be equivalent for human welfare, which indirectly suggests a weakness in just having a target set by a future date. To evaluate the consequences for human welfare, however, requires us to think about the climate-economy interactions, to which we turn in Section 25.4 below. A final point regards the linearity: a constant CCR also means a ruling out of non-linearities such as tipping points in the global climate.

25.3.4 The fossil energy supply

Against the background of a carbon budget, the question of how much global warming that can be expected depends not only on the CCR, but also on how large the stock of fossil fuel actually is. Here, there is an important distinction between *reserves* and *resources*. Using the definition from the BP (2021), **total proven reserves** is generally taken to be those quantities that geological and engineering information indicates with reasonable certainty can be recovered in the future from known reservoirs under existing economic and operating conditions. **Resources** includes the amount of a geologic commodity that exists in both discovered and undiscovered deposits. Hence, it is by definition a "best guess" that includes quantities that would not be profitable to extract today but that could become profitable in the future.

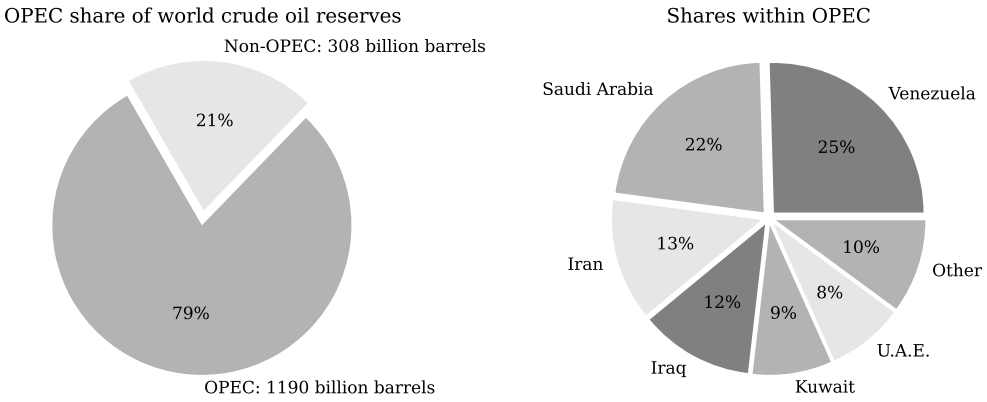


Figure 25.4: Oil Supply.

Source: OPEC (2019).

Starting with oil, world proven crude oil reserves are displayed in Figure 25.4. The figure shows that OPEC accounts for almost 80 % of world crude reserves with their 1,190 billion barrels, whereas the non-OPEC reserves consist of 308 billion barrels.

It is also useful to make a distinction between "conventional" and "unconventional" oil. The key difference is that conventional oil can be extracted with traditional drilling methods: it is liquid at atmospheric temperatures and therefore flows without additional stimulus. Unconventional oil requires more advanced extraction methods because it does not flow on its on. Marginal extraction costs are therefore (very) low for conventional oil: only a fraction of the market price. The approximation of marginal cost to 0 in the section

above on Hotelling rents precisely referred to conventional oil. For unconventional oil, the marginal extraction cost is substantially higher and much closer to (or above) the market price. For these reasons, conventional oil is very profitable, whereas unconventional oil is not. As we will see below, all the conventional oil can therefore be expected to be used up.

How much global warming would result if we were to burn all the reserves of conventional oil? It is somewhat hard to assess exactly how much of the oil that is conventional, but it seems reasonable to assume that most of the middle-east reserves are conventional. To compute the level of warming, we then make use of the approximately constant CCR: each 1,000 $GtCO_2$ of cumulative CO_2 emissions is likely to cause a 0.27–0.63 °C increase in global surface temperature. Converting from barrels to $GtCO_2$, we note that one barrel equals 1/7.33 ton, that the carbon content in oil is 85%, and that one ton of carbon equals 3.67 ton of CO_2. We thus have that 1,190 billion barrels equals $(1,190/7.33)\cdot 0.85 \times 3.67 = 500 GtCO_2$. Using constant CCR, this delivers 0.13–0.33°C of warming.

Turning to coal, total proven reserves are 1,074 GtC, according to BP (2021). With the carbon content of coal being about 70 percent of its weight. This corresponds to $1,074 \cdot 0.70 \cdot 3.67 = 2,759$ $GtCO_2$, which if all burnt results in an increase in the temperature of between 0.75°C and 1.73°C. It is important to point out, however, that proven coal reserves are likely underestimation: coal is not particularly profitable, giving only weak incentives to search for it.

Total proven reserves of natural gas, finally, is 190 trillion cubic meters, again according to BP (2021). Using the conversion factor 0.511 kgC/cubic meter, we see that burning all the gas results in less than a tenth of a degree of warming.

What about resources? The numbers are much more uncertain and depend on how likely it is that some resources become profitable to extract. Estimates of resources (and reserves) are provided in Table 25.1.

Table 25.1: Reserves and resources in GtC

	Reserves	Resources (Rogner)	Resources (BGR)
Oil (conv+unconv)	173	$\approx 400 - 2,200$	≈500
Coal	1,074	≈6,200	≈11,500
Gas	100	≈350	≈365

Sources: Rogner (1997), BGR (2020). BGR is the Federal Institute for Geosciences.

As can be seen the resources of coal are an order of magnitude larger than the reserves; clearly, if this whole amount of coal were burnt, the global temperature would rise by at least 10°C, again using the approximate CCR.

25.3.5 Damages

In October of 2007, the *Stern Review on the Economics of Climate Change*—a report commissioned by the British government—was released and it had great impact in the policy-making world, arguably precisely because of its discussions of how climate change would "damage" the economy. It pulled together evidence and made a quantitative assessment of

the economic value lost from warming, resulting in strikingly large numbers. This finding, the review argued, called for strong, early action by governments to halt global warming. In the present section we discuss the effects of climate change on human welfare, drawing on a growing literature, which includes touching on the findings in the Stern Review.

From an economic perspective, as the Review and the earlier work by others such as William Nordhaus had pointed out, the effects of warming on human welfare are to be regarded as classical externalities: consequences of economic behavior (emissions in the context of burning fossil fuel) for human welfare that are not priced by the market. An approach for policy, proposed in the Stern Review, by Nordhaus, as well as by many other influential economists, is therefore to aggregate all the externalities (measured in dollar terms), and tax emissions by that amount: the so-called Pigou tax.[10] Moreover, in the case of climate externalities, though it was not emphasized in the previous sections, because carbon dioxide emissions spread very quickly around the globe, the total externality is the same independently of where the emission takes place. Thus, an application of a Pigou tax in this case means that the tax per unit of emission should be the same everywhere. There is no universal agreement on a Pigou tax being the best policy, and there is even less agreement that a tax, if applied, should be set at the same level in all countries. Regardless, the measurement of damages is still a very important part of the climate-economy field.

Economists have approached damage measurement in two different, and complementary, ways. One is the *bottom-up* approach, whereby all kinds of damages are identified, with a variety of econometric methods, assigned dollar values using market prices, and then added up. Clearly, there is a space/country dimension here, so the endeavor of covering all damages is extremely ambitious. Another approach is the *top-down* method, which focuses on relating aggregates—such as production, mortality, or migration—to observed changes in the climate over time or to climate differences across space. We now go through some results from this (rapidly growing) literature.

The bottom-up approach

The idea here is to measure damages of all kinds imaginable (i.e., broadly defined), assign them (dollar) values, and then add up over a given time period such as a year, thus generating a flow value. Moreover, the goal is to estimate the damage effects of each additional degree of global warming (along with the broader notions of climate change, such as changes in precipitation patterns, that warming entails), thus obtaining a *damage function*, i.e., a mapping from global temperature to a dollar value. Damages functions can be constructed separately for different countries or regions, for different sectors of production, and so on. Damages will, moreover, depend on time or economic state variables. For example, if warming affects output, then economic growth can make damages larger if they are proportional to output: there is more to lose with higher output. If people's health is negatively affected too, then the number of people factors in, so that population growth too will affect the measure of damages. Clearly, one cannot literally include everything in a global damage measure, at all

[10] The so-called Pigou Club, formed in 2006 by N. Gregory Mankiw with the general aim to call for governments to correct externalities with taxes and subsidies, contains a long list of very well-known economists. Arthur Pigou was a British economist who introduced the externality concept and proposed this policy; see Pigou (1920).

points in time and space and for all contingencies, so for the approach to be implementable in practice it is important to make attempts generalize from specific studies.

Nordhaus pursued a bottom-up approach, thus looking separately at damage effects via agriculture, sea-level rise, health, non-market amenity impacts, etc., but in adding up found limited quantitative impacts. He included the potential for catastrophes, with probability and severity assessments collected in a survey among experts rather by direct measurement and still arrived at a damage function with very modest effects: global warming of 2 degrees—the Paris agreement now aims at a maximum of 1.5 degrees—would only generate a flow damage value of less than a percent of global GDP. Nordhaus argued that the damage function was likely convex, though the degree of convexity is hard to assess.[11] He also formulated separate damage functions for world regions but associated great uncertainty here as well.

Later on the European Commission started the PESETA project, which has studied 11 damages categories, including heat waves, windstorms, droughts, flooding, wildfires, agriculture and energy supply, across the EU. PESETA IV reports aggregate EU damages that are nontrivial (on the same order of magnitude as those found by Nordhaus) with significant dispersion across different parts of the EU: the north is even estimated to gain from warming, but the Mediterranean economies lose significantly.

The number of studies with microeconometric damage estimates is growing rapidly, which is reassuring. We are, however, still far from a comprehensive understanding of the nature of the effects of warming on people and on our economies.

The top-down approach

The top-down approach starts from aggregate information, typically on production/GDP or other economic aggregates, such as health indicators and mortality, and then tries to relate these figures to climate variables. There are multiple econometric challenges in interpreting the typical regressions of this sort, but they have nevertheless played an important role in the literature on damages. The end results, moreover, are "reduced forms", i.e., not pinpointing through what channels warming affects output or mortality.

One approach has been to use time-series analysis where climate change is represented by changes in weather (temperature) over time. Here one challenge is that "climate" is the "distribution of weather", so that short-run changes in temperature (and the economic reactions to them) cannot necessarily be interpreted as climate change. Dell, Jones, and Olken (2014) find, based on a time series study of 125 countries over the period 1950–2005, that a one degree higher temperature leads to a one percent fall in output growth among poor countries but no effect for rich countries, and no level effects more generally. Kahn, Mohaddes, Ng, Pesaran, Raissi, and Yang (2019) use data from 174 countries over the period 1960–2014 and also find sizable effects on growth, e.g., that a four degree warmer global average temperature would reduce global GDP by 0.04 percent per year, accumulating to a reduction in global GDP in 2100 by 7%. A recent study on mortality (Carleton, Jina, Delgado, Greenstone, Houser, Hsiang, Hultgren, Kopp, McCusker, Nath, Rising, Rode, Seo, Viaene, Yuan, and Zhang, 2022) finds, in an extreme scenario with a 5-degree increase in

[11] Nordhaus (1994).

global average temperature by 2100, that the total global cost of mortality (and adaptation to avoid it) amounts to 3.2% of GDP.

Another approach is to rely on cross-sectional data to glean the effects of climate on aggregates (regions or countries). This approach is related to Hall and Jones (1999), which accounts for differences in output by measuring the distance to the equator (which is closely related to average temperature), but it is actually preceded by Mendelsohn, Nordhaus, and Shaw (1994), which tries to explain agricultural output in a similar manner, with some attempts to hold constant institutions. Dell et al. (2014) also highlight the remarkable negative correlation between output and average temperature. Looking at regions at an even higher level of resolution, Cruz and Rossi-Hansberg (2022) and Krusell and Smith (2022) find a U-shape in the relation between average temperature and productivity, with a "sweet spot" around 11 degrees Celsius (measured as an annual average). Clearly, these cross-sectional estimates carry information about the effects of the climate on output but, equally clearly, there are confounding factors.

Summaries and modeling

We now summarize the damage estimates in the literature by looking at two meta studies and then comment on how damages are modeled in the IAMs.

Meta studies We report two meta studies relating global warming to losses in global GDP in Figure 25.5. Both of these studies draw on a large number of different contributions using a variety of methods. As the figure shows, the costs of one or two degrees of warming are on average quite low but there is convexity as a function of temperature.

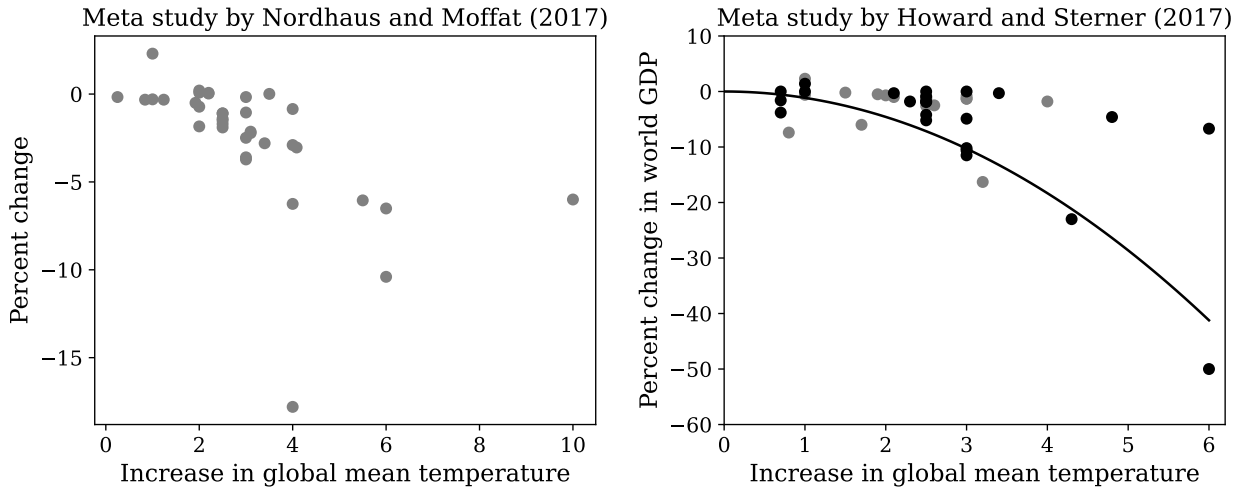


Figure 25.5: Meta studies of damages. Left: study by Nordhaus and Moffat (2017). Right: study by Howard and Sterner (2017) with black dots showing non-duplicate studies and the regression line preferred by those authors.

Damages: modeling the function(s) How can the economic damages be modeled? Nordhaus pioneered an approach where all damages are contained in a TFP factor. Nordhaus

and Sztorc (2013) specifies damages, as a share of GDP, at a temperature T degrees above the pre-industrial level as given by

$$D(T) = 1 - \frac{1}{1 + 0.00267T^2},$$

where the coefficient in front of T^2. This formulation can be approximated locally around 0 by $0.024\left(\frac{T}{3}\right)^2$, a convex function indicating that at three degrees of warming, 2.4 percent of output is lost.[12]. Nordhaus also considered including higher order polynomials; Weitzman argued for an order of 6.8. The higher orders, however, are not based on the use of data but rather as examples.

It is of course also possible to introduce damages in other parts of one's integrated assessment model. Endogenous life-expectancy fits naturally by introducing variable population size and a value of life, one can have endogenous depreciation by taking capital destruction (due to, say, flooding) into account, and so on. One common approach is to simply have damages additively in utility—we describe a simple model of that sort below—because in such a case there is no feedback from damages to economic activity, which makes the model simpler to solve.

An interesting case is obtained by mapping the carbon dioxide concentration in the atmosphere, S above, directly into percentage TFP losses. In particular, Golosov et al. (2014) show that the maps from S to T, by the use of the Arrhenius law described in Section 25.3, and then further from T to D according to Nordhaus's estimated function, can be very well approximated by a TFP factor $e^{-\gamma S}$. I.e., if output before damages is multiplied by this expression we obtain a surprisingly good approximation (for a specific numerical γ value) to the original formulation. This, we shall also see below, turns out to be another useful case for applied use.

Damages: adaptation In response to global warming and damages therefrom, it is possible to respond in a variety of ways to lessen the damages. This phenomenon is referred to as adaptation. For example, air conditioning can be installed in response to heat increases, walls can be built to protect against flooding, and so on. Reduced-form damage assessments thus should be interpreted as incorporating adaptation, whereas at least some microeconometric damage estimates simply do not factor in the possibility of adaptation.

There are reasons to think that damages net of adaptation are less convex than damages not taking adaptation into account. Consider the following example: total damages, in output units, are

$$\frac{(T-a)^x}{x} + a,$$

where a is adaptation. This formulation captures a possible convexity—with $x > 1$—and the assumption that adaptation costs scale linearly: a wall twice as high costs twice as much. Minimizing this expression with respect to a delivers adaptation of $T-1$ (so long as $T \geq 1$). The reduced-form damage is then, again for $T > 1$, $D(T) = T + \frac{1-x}{x}$, i.e., the damage function is linear, regardless of the value of x!

[12]Globally, the damage function is not convex: it asymptotes to 1 (all of output is lost) as T goes to infinity.

Damages: discounting Finally, we comment very briefly on a contentious issue: how to add up damages that occur over time. In a typical macroeconomic model based on microeconomic foundations, welfare for each agent is simply measured by the utility function of that agent, using a revealed-preference argument. So if we have measures of utility function parameters, the revealed-preference approach would be straightforward to apply. In a representative-agent model with standard discounting, we would use an estimate of β, the discount factor, to add up utils accruing at different points in time: a utility flow occurring t periods into the future is then discounted by β^t. We also explained in our early chapters that estimates of β come from observations on the average long-run (riskless) real interest rate and consumption growth together with the Euler equation and a measure for consumption curvature. Thus, a β around 0.985 is standard for an annual time period. This approach is followed by Nordhaus in his work.

In the Stern Review, an argument is put forth that a much higher β, e.g., 0.999, is appropriate. The idea is thus to "assign a higher weight to future generations", which may not be appropriately taken into account using the revealed-preference approach above. The choice of β matters greatly on a quantitative level, since global warming is (largely) irreversible and damages long-lasting. In our view, the choice is more a matter of philosophy, not economics. It deserves to be pointed out, however, that if one disagrees with market discounting (for philosophical reasons), there may also be reason to revisit the assessments of how markets make investment decisions more generally, potentially thus implying a need for large subsidies to saving and investment.

25.4 Integrated assessment models

We are now ready to put together an integrated assessment model. The quantitative model we will use is, as advertised, a neoclassical growth model in its core, but we will actually begin with a static model, as it can provide us with significant insight, including how to think about climate policy.

25.4.1 A static one-region model

The static model is a very simple static version of the macroeconomic models earlier in the book. It might, at first thought, seem strange to think of climate change in a static setting since change by definition is non-static. However, this setting has proven useful in providing intuition for much more complex models.[13] The model has three sectors and all sectors are perfectible competitive. The final-goods sector uses capital, labor, and energy services to produce a final good. Energy services is a new element relative to earlier chapters. We assume a very simple structure whereby these energy services are produced one-for-one from coal. Coal, finally, is produced at a constant marginal cost, labeled p, in terms of the output good.[14] Coal is thus for simplicity not regarded as a resource in finite supply here. Our

[13] See, for instance, Hassler, Krusell, and Olovsson (2021b).

[14] Recall from earlier in the book that this is equivalent to regarding coal as being produced by the same basic inputs, and with the same production technology isoquants, as for final goods, only at a relative TFP of $1/p$.

dynamic model below will, however, consider fossil fuel in finite supply.

The production, Y, of the final good is given by

$$Y \equiv F(K, L, E) = AK^{\alpha}L^{1-\alpha-\nu}E^{\nu}, \tag{25.11}$$

where A is total factor productivity, K is the capital stock, L is labor, and E energy services. Unitary elasticity between inputs is a good approximation in the longer run (and the static model is meant to capture a longer run, e.g., 100 years) but not in the shorter run, where the elasticity is much lower between energy and the other inputs. In Section 25.5.3 we discuss this issue in detail. Without loss of generality, we normalize L and K to unity, which implies that output can be written as

$$Y \equiv f(1, 1, E) = AE^{\nu}. \tag{25.12}$$

In a decentralized allocation, the energy-service provider buys the energy input at marginal cost from the fuel producer. The profit-maximizing problem in the final-goods sector is then given by

$$\pi_e = \max_{E} AE^{\nu} - (p + \tau)E, \tag{25.13}$$

where τ is a tax on coal use. We will, below, refer to (25.13) as the private economy since this is the profit-maximization problem for a representative firm in the market economy.

This simple IAM assumes a direct link between coal use and damages. As explained above, there exists a mapping from E to S (see equation (25.10)) and then from S to T (see equations (25.4)-(25.5)). Here, these relationships are summarized in a way that implies that the damage function is quadratic in coal use:

$$D\left(e_1\right) = \gamma E^2. \tag{25.14}$$

The above function captures the typical property in climate models that was described above, i.e., a proportional relation between accumulated emissions and the temperature. Combining this property with the assumption that damages are moderately convex in temperature yields (25.14). Damages are, moreover, assumed to be separable from output, which implies that consumption is given by

$$C = AE^{\nu} - pE - \gamma E^2. \tag{25.15}$$

Let us now consider the equilibrium in the market economy. Because damages are additively separable, the temperature and the damages do not appear in the problem for the firm, i.e., equation (25.13). The first-order condition for this problem can then be written as

$$\nu AE^{\nu-1} = p + \tau. \tag{25.16}$$

The only unknown in the above equation is E so this variable is uniquely pinned down by(25.16); it is given by $E = ((\nu A)/(p + \tau))^{\frac{1}{1-\nu}}$. Note that there is direct negative relationship between E and τ so coal use is decreasing in the coal tax. Note also that, with mappings between E and S and T specified, it is possible to compute the temperature increases that result from different tax rates in the static period.

Pigou taxation

Turning to the social-planning problem, the objective is given by

$$\max_E AE^\nu - pE - \gamma E^2, \tag{25.17}$$

and the first-order condition can be written as

$$\nu A E^{\nu-1} = p + 2\gamma E. \tag{25.18}$$

Clearly (25.16) and (25.18) differ if $\gamma \neq 0$ and $\tau = 0$. In this case, the market economy will be inefficient and emissions will be too high relative to the efficient level. Even though we cannot solve for τ in closed form, it is clear that implementing the first best in the market economy implies setting $\tau = 2\gamma E^*$, where E^* is the solution to (25.17). This is Pigou's recipe: tax the activity that is causing a social cost and set the tax equal to the marginal externality damage that the activity is causing (Pigou, 1920). At the social optimum in this economy, the marginal damage is equal to $2\gamma E^*$.

For the private economy, i.e., ignoring the externality entirely, $\tau = 0$ constitutes a maximum. Moreover, this objective function is smoothly mountain-shaped, and concave, around a zero tax. It also implies that the costs for the private economy of modestly raising the global carbon tax above zero are negligible. The benefits for overall welfare, however, can potentially be very large if γ is large. In other words, even a modest tax can be very beneficial in alleviating the negative consequences of global warming.[15]

Regulation: prices vs. quantities

In practice, it seems that quantity restrictions have been preferred by politicians over affecting prices by the use of taxes. In 2005, the European Union implemented a cap-and-trade system—the EU Emission Trading System (EU ETS)—which is a quantity restriction where firms also can trade in emission rights. This raises the question of how effective quantity restrictions are relative to Pigou taxation.

To answer that question, consider a social planner that implements a quantity restriction. The regulator then decides on a quantity of coal that cannot be exceeded by the market. To ensure that the restriction is honored, anyone that wants to burn a unit of coal (and thus emit carbon dioxide into the atmosphere) also has to hand in an emission right to the regulator for doing so. A quantity restriction therefore requires the regulator to take a stance on how the emission rights are to be allocated. This can be carried out in many ways: the emission rights can, for instance, be handed out to one or several agents for free. Alternatively, the government can sell them in an auction format. Here, we simply assume that the emission rights are handed out for free to the producer of coal. This firm then sell these emission rights to the final-output firm (which, in this economy, is the only one that buys coal).

Every time that the final-good firm now wants to buy a unit of coal, it also has to buy an emission right from the coal producer. The total cost of buying a unit of coal is then $p+\lambda$, where λ is the price of the emission right. Note that the usage of a quantity restriction

[15] See Hassler, Krusell, and Olovsson (2023).

transforms the problem from one where, in principle, an infinite amount of coal can be supplied to one where coal effectively becomes a resource in finite supply.

Consider first the case where the quantity restriction requires $E \leq E^*$, i.e., we target the optimal level. The first-order condition for the final-output firm reads

$$\nu A E^{\nu-1} = p + \lambda. \tag{25.19}$$

Will E^* be an equilibrium outcome and, if so, what will the equilibrium price of emission rights be? It is easy to see that, by construction, $\lambda = \tau = 2\gamma E^*$ satisfies the first-order condition and, hence, is an equilibrium. At this quantity, λ has to be positive, since otherwise more E would be purchased than what is available. A lower E is not possible: then all emission rights would not be sold, but that can only occur if $\lambda = 0$, which clearly does not satisfy the first-order condition. The equilibrium is unique. Hence, the price and quantity regulations are equivalent.[16]

What is the price of an emission right if the quantity restriction instead selects another quantity? If the quantity is higher than that of the laissez-faire equilibrium, the restriction will not bind and λ will be zero. If it is lower than that, the restriction will bind and one reads off λ from the firm's first-order condition.

Finally, let us briefly consider how the efficient prices and quantities are affected by firm heterogeneity. For simplicity, consider the case with two final-output firms, labeled 1 and 2, that differ in their technologies with respect to energy intensity (ν_1 vs. ν_2). The damage function then reads $D(E_1 + E_2) = \gamma(E_1 + E_2)^2$. The planning problem reads $\max_{E_1,E_2} AE_1^{\nu_1} + AE_2^{\nu_2} - \gamma(E_1 + E_2)^2$ and the first-order condition for firm $i \in \{1, 2\}$ is then $\nu_i A E_i^{\nu_i - 1} = 2\gamma(E_1 + E_2)$, which solves uniquely for $\{E_1^*, E_2^*\}$. A carbon tax satisfying $\tau = 2\gamma(E_1^* + E_2^*)$ would deliver the optimal outcome. What about a quantity regulation: would the government need to assign two values for quantities, one for each technology? That would be a possible approach—in theory, given if the government knows all the details of how firms differ—but it is not necessary: a mere restriction for total use not to exceed $E_1^* + E_2^*$ would also work (the equilibrium price of emission rights would again equal the optimal tax). In particular, it would ensure an optimal allocation of energy across firms, since they all face the same price (or tax): $\nu_1 A E_1^{\nu_1 - 1} = \nu_2 A E_2^{\nu_2 - 1}$. This is exactly how the EU system works.

25.4.2 A fully dynamic integrated assessment model

The static one-region IAM described in the previous section can straightforwardly be extended to a fully dynamic model which is also richer in other ways. The model that we will formulate here builds on that described in Golosov et al. (2014) and it incorporates 2 regions. Region 1 is inhabited by a representative oil producer that is endowed with a finite amount of conventional oil that it extracts and sells to the other region in a competitive manner. We refer to this region as the *oil producer*. The other region does not have any endowments of conventional oil but, given that they need oil, import it from the oil producer and pay for this import with a common final good. This region is referred to as the *oil consumer*. Each

[16] Weitzman (1974) shows that random fundamentals can, under some conditions, break the equivalence.

region is inhabited by a representative consumer with preferences given by

$$\sum_{t=0}^{\infty} \beta^t \log(C_t). \tag{25.20}$$

The law of motion for the atmospheric excess stock of carbon S_t is given by (25.9). The climate is then affected by the concentration of CO_2 in the atmosphere via the greenhouse effect. Here, we employ the direct, exponential relationship between atmospheric carbon concentration S and damages discussed in Section 25.3.5, but modify it slightly so that the production function reads

$$Y_t = \underbrace{\exp\left(z_t - \gamma S_t\right)}_{\equiv A_t} L_t^{1-\alpha-\nu} K_t^{\alpha} E_t^{\nu}, \tag{25.21}$$

where z is a potentially stochastic productivity trend and γ is a parameter that determines how climate-related damages depend on the level of the atmospheric CO_2 concentration. The climate system, i.e., the temperatures follow the energy budget model in DICE and RICE and is thus given by (25.4) and (25.5).

Energy services in the oil-consuming region are produced and supplied by a local competitive representative firm that combines different energy sources as inputs. There are three available energy sources: conventional oil, coal, and a green energy source.[17] All energy sources except conventional oil are locally produced by competitive firms that are using a production technology that is linear in the final good, as in the static model.[18] Specifically, $p_{\kappa,t}$ units of the final good are required to produce $e_{\kappa,t}$ units of the energy source $\kappa \in \{c, g\}$ in period t, where c and g refer to coal and green, respectively. The total amount of energy services is an aggregate of the energy inputs

$$E_t = \left[\lambda_o e_{o,t}^{\rho} + \lambda_c e_{c,t}^{\rho} + \lambda_g e_{g,t}^{\rho}\right]^{\frac{1}{\rho}}, \tag{25.22}$$

where ρ determines the elasticity of substitution, and $\lambda_o + \lambda_c + \lambda_g = 1$.

Final goods are either consumed, invested or used for energy production, and capital is assumed to fully depreciate between periods. This assumption is not too problematic given that a period will be calibrated to be ten years. The resource constraint for the final good is then given by

$$C_t + K_{t+1} = A_t L_t^{1-\alpha-\nu} K_t^{\alpha} E_t^{\nu} - p_{o,t} e_{o,t} - p_{c,t} e_{c,t} - p_{g,t} e_{g,t}, \tag{25.23}$$

where $p_{o,t}$ denotes the world market price for conventional oil. Total emissions in period t are given by

$$M_t = e_{o,t} + e_{c,t}, \tag{25.24}$$

where this simple summation reflects the fact that the fossil energy sources all are measured in carbon units.

[17] It is straightforward to include additional energy sources. Hassler et al. (2021b) also incorporate hydraulic fracturing or "fracking".

[18] As before, we abstract from the finiteness of coal supply as the amount of coal is so large that the resource restriction is unlikely to bind, unless we are willing to accept global warming at the two-digit Celsius level.

The oil producer extracts oil without any extraction costs, and the total stock of oil in the ground at time t has size R_t.[19] With extraction in period t given by $e_{o,t}$, the law of motion for the stock of oil is given by

$$R_{t+1} = R_t - e_{o,t}, \tag{25.25}$$

subject to

$$R_t \geq 0, \forall t.$$

The government, finally, simply sets a carbon tax and then recycle the proceeds back to the household within the period in the form of a negative income tax rate, Γ_t. The government budget constraint is then given by

$$\Gamma_t \left(w_t L_t + r_t K_t\right) = \tau_t(e_{o,t} + e_{c,t}).$$

Households supply labor inelastically and maximize (25.20) subject to the budget constraint

$$C_t + K_{t+1} = (1 + \Gamma_t)\left(w_t L_t + r_t K_t\right) = (1 + \Gamma_t)\,\widehat{Y}_t,$$

where $\widehat{Y}_t \equiv (1 - \nu)Y_t$ is output net of energy expenses.[20]

Equilibrium

We now describe how to compute the market equilibrium. The problem of the oil supplier has a very simple solution. The budget constraint of this agent can be written $C_{o,t} = p_{o,t}e_{o,t}$, where C_o is the oil supplier's consumption. The maximization of $\sum_{t=0}^{\infty}\beta^t \log(C_{o,t})$ subject to this constraint and (25.25) straightforwardly delivers $e_{o,t} = (1-\beta)R_t$ for all t. We discuss "optimal extraction" much more in Section 25.5 below.

The problem for the energy-service provider is to derive the demand functions for all fuels; it reads

$$\min_{e_{o,t},e_{c,t},e_{g,t}} p_{o,t}e_{o,t} + p_{c,t}e_{c,t} + p_{g,t}e_{g,t} - P_t\left(\left[\lambda_o e_{o,t}^{\rho} + \lambda_c e_{c,t}^{\rho} + \lambda_g e_{g,t}^{\rho}\right]^{\frac{1}{\rho}} - E_t\right),$$

where P_t is both the Lagrange multiplier on the constraint and the price index of energy services. We know from Chapter 6 that cost minimization for the energy-service provider delivers that this price index is a CES function of the different input costs. The price of energy services can then be shown to be given by

$$P_t = \left(\lambda_o^{\frac{1}{1-\rho}}\hat{p}_{o,t}^{\frac{\rho}{\rho-1}} + \lambda_c^{\frac{1}{1-\rho}}\hat{p}_{c,t}^{\frac{\rho}{\rho-1}} + \lambda_g^{\frac{1}{1-\rho}}\hat{p}_{g,t}^{\frac{\rho}{\rho-1}}\right)^{\frac{\rho-1}{\rho}}. \tag{25.26}$$

The first-order conditions with respect to $e_{\kappa,t}$ then yield

$$e_{\kappa,t} = E_t\left(\frac{P_t\lambda_\kappa}{p_{\kappa,t}}\right)^{\frac{1}{1-\rho}}, \text{ for } \kappa \in \{o, c, g\}. \tag{25.27}$$

[19] The approximation that oil is costless to produce is good for conventional oil/natural gas but not for non-conventional resources such as fracking or deep-sea oil.

[20] The assumption of constant returns to scale implies that profits are zero in equilibrium, and that energy expenses account for the share ν of output.

Producers of the final good maximize profits taking P_t as given, implying that $P_t = \nu A_t L_t^{1-\alpha-\nu} K^{\alpha} E_t^{\nu-1}$, which can be rearranged to solve for energy service use:

$$E_t = \left(\nu \frac{A_t L_t^{1-\alpha-\nu} K^{\alpha}}{P_t} \right)^{\frac{1}{1-\nu}}. \tag{25.28}$$

Similarly, wages and the interest rate will be given by $w_t = (1-\alpha-\nu)Y_t/L_t$ and $r_t = \alpha Y_t/K_t$, respectively. The household's Euler equation reads

$$\frac{C_{t+1}}{C_t} = \beta \left(1 + \Gamma_t\right) r_{t+1}. \tag{25.29}$$

Defining the savings rate out of net output as $s_t = (\widehat{Y}_t - C_t)/\widehat{Y}_t$, we obtain $C_t = (1-s_t)(1+\Gamma_t)\widehat{Y}_t$ and $K_{t+1} = s_t(1+\Gamma_t)\widehat{Y}_t$. Inserting these expressions and $r_{t+1} = \alpha Y_{t+1}/K_{t+1}$ into (25.29) yields

$$\frac{1+s_{t+1}}{1+s_t} = \frac{\alpha\beta}{s_t(1-\nu)}.$$

The difference equation in s_t defined by the above expression only has one non-explosive solution: $s_t = (\alpha\beta)/(1-\nu) \equiv s, \forall t$. The savings rate is thus constant and it defines optimal household behavior.

The equilibrium can now be defined by equation (25.21)–(25.28), $s_t = (\alpha\beta)/(1-\nu)$, and where the state variables evolve according to $K_{t+1} = \frac{\alpha\beta}{1-\nu}\left(1+\Gamma_t\right)\hat{Y}_t$, $R_{t+1} = \beta R_t$, and $S_t = \sum_{v=0}^{t}(1-d_{t-v})\sum_i M_t$.

Note that the allocation is determined sequentially without any forward-looking terms. This is due to the combination of logarithmic utility, Cobb-Douglas production, full depreciation, and the way that tax revenues are rebated. Moreover, except for the world market price of oil all equilibrium conditions have closed-form solutions. Hence, when it comes to solving the model, finding the equilibrium is only a matter of finding the equilibrium oil price in each period. This is easy since the supply in each period is predetermined at $(1-\beta)\,R_t$. As a result, the model typically solves in under a second also when it is expanded to include many regions.

Turning to the social planning problem, it is given by

$$\max_{\{C_t, K_{t+1}, e_{o,t}, e_{c,t}, e_{g,t}, R_{t+1}\}_{t=0}^{\infty}} \mathbb{E}_t \sum_{t=0}^{\infty} \beta^t \log(C_t)$$

subject to (25.9), (25.21), (25.23), and (25.25). In a more general version of this model, it is possible to define the marginal damage externality caused by energy source κ, Λ_t^{κ}, to be given by

$$\Lambda_t^{\kappa} = \mathbb{E}_t \sum_{j=0}^{\infty} \beta^j \frac{U'(C_{t+j})}{U'(C_t)} \frac{\partial F_{t+j}}{\partial S_{t+j}} \frac{\partial S_{t+j}}{\partial e_{\kappa,t}}, \tag{25.30}$$

where U is a more general utility function and F is a more general production function for final goods. The intuition is clear and very robust: a unit emitted today gives $\frac{\partial S_{t+j}}{\partial e_{\kappa,t}}$ units of increased atmospheric carbon concentration j periods later, which in turn changes output

by $\frac{\partial F_{t+j}}{\partial S_{t+j}}$ per unit. Standard discounting and adding up of these future output effects then gives us the formula.

With the functional form of the production function that we have considered here, we have

$$\frac{\partial F_t}{\partial S_t}\frac{1}{Y_t} = \frac{\partial\left(e^{-\gamma S_t}Y_t\right)}{\partial S_t}\frac{1}{e^{-\gamma S_t}Y_t} = -\gamma. \tag{25.31}$$

The implication of the above equation is that every additional unit of carbon in the atmosphere yields a constant percentage reduction in GDP. In other words, the "marginal damage flow" is independent of both GDP and the CO_2 concentration. Moreover, the marginal utility of consumption, $U'(C_t)$, just becomes $1/C_t$ under logarithmic utility. Finally, the derivative $\frac{\partial S_{t+j}}{\partial e_{\kappa,t}}$ is straightforwardly computed from (25.9). Combining these expressions, we can now write the marginal damage, which according to the Pigou principle also is the optimal tax on carbon, in the following way.

$$\tau_t = Y_t\left[\mathbb{E}_t\sum_{j=0}^{\infty}\beta^j\gamma(1-d_j)\right]. \tag{25.32}$$

The expression for the optimal tax rate is surprisingly simple. The marginal externality cost of emissions as a proportion of GDP inherits the time path of GDP. The tax rate then depends on three "d:s". The first one is *depreciation* (d), i.e., how long the marginal unit of emitted carbon dioxide is staying in the atmosphere. Since a unit will generate more damages the longer it stays in the atmosphere, the higher the tax will be. The second d is the *damage* as measured by γ. Naturally, the tax is increasing in γ since this is a direct measure of the damages that are generated by the emission. The third d, finally, is the *discount factor*, β, i.e., the weight that is put on the future. Clearly, the tax is also higher if a higher weight is put on the future. What is remarkable about the formula is that it does not contain any other parameters, for example involving how fossil fuel is produced. It is, moreover, a reasonable benchmark case since logarithmic utility and Cobb-Douglas utility are standard assumptions in the macroeconomic literature; $\delta = 1$ is a decent approximation if the time period is long enough.

Simulations

We are now ready to show some output from the model. We are particularly interested in evaluating the effect on the global temperature of different carbon taxes. We consider three levels for the tax: 0, an optimal tax computed from (25.32), and a moderate carbon tax that we take to be one third of the optimal tax. To evaluate these policies, we first need to calibrate the model; this is carried out in the appendix and follows Golosov et al. (2014) and Hassler et al. (2021b). The model has many parameters but almost all of them can be set to match standard long-run economic facts or facts from the natural-science material presented above. The observed prices and shares of different energy sources can be used to select the energy-related parameters; the elasticity of substitution between different energy sources is hard to calibrate and we set it to 2, i.e., a significantly higher substitution elasticity than for Cobb-Douglas. We use Nordhaus's benchmark damage estimate.

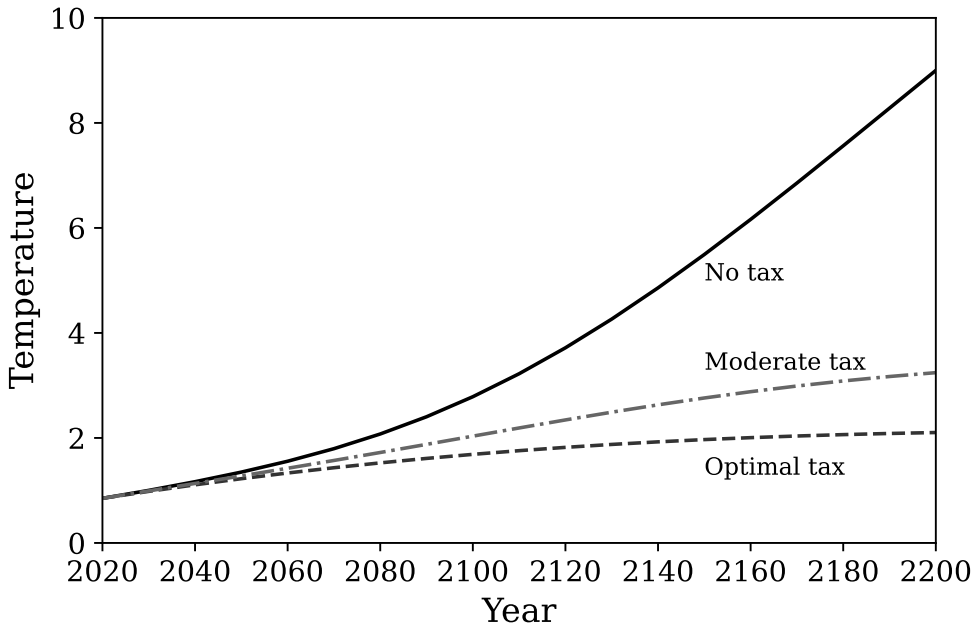


Figure 25.6: The effect om the global temperature of different carbon taxes.

Notes: The optimal tax is set to U.S. \$25 per ton CO_2 and the moderate tax is U.S. \$8 per ton CO_2.

Given the calibration, the optimal tax is about U.S. \$25 per ton CO_2 (or U.S. \$90 per ton carbon). This relatively low number is clearly sensitive to assumptions about the three "d"s in equation (25.32).[21] The moderate tax is then only about U.S. \$8 per ton CO_2. The results associated with these different carbon taxes are presented in Figure 25.6. Note the very large difference in temperature increase between the laissez faire with no tax and the case with the optimal tax. Without any climate policy, the global temperature will have increased by as much as 9°C by the year 2200. In contrast, with a global optimal tax in place, the increase will be contained at about two degrees Celsius at the end of the considered period. This shows the strength and potency of the carbon tax as a policy for mitigating global warming. Note also that even a very low carbon tax (i.e., the moderate tax) can go a long way in making sure that the temperature increase is not too large.

25.5 Natural resources in finite supply

We now turn back to the broader issue of finite resources. When a resource is in finite supply, a basic issue is the normative one of how to manage the resource properly: at what rate should the resource be used over time? Another question is how markets, left to their own devices, will actually manage the resource. An associated question is how the resource price will evolve over time. From the perspective of climate change, it also becomes important to think about how the economy can manage without fossil fuels: can we economize on energy

[21] Again, see Golosov et al. (2014) for a detailed discussion.

use by altering the way we produce and consume? The goods and services that are very energy-intense in production can perhaps be replaced by other products. Alternatively, we can explore other sources of energy. These issues will all be discussed in this chapter. We begin with a core theory that rests on Harald Hotelling's seminal contribution from 1931.[22]

25.5.1 Some data

Before we dive in to the theory, it is useful to look at some data. Figure 25.7 shows U.S. data on the uses of different renewable as well as non-renewable energy sources over a long time span. We see clear upward trends, while there are also (potential) peaks in use for oil and coal.

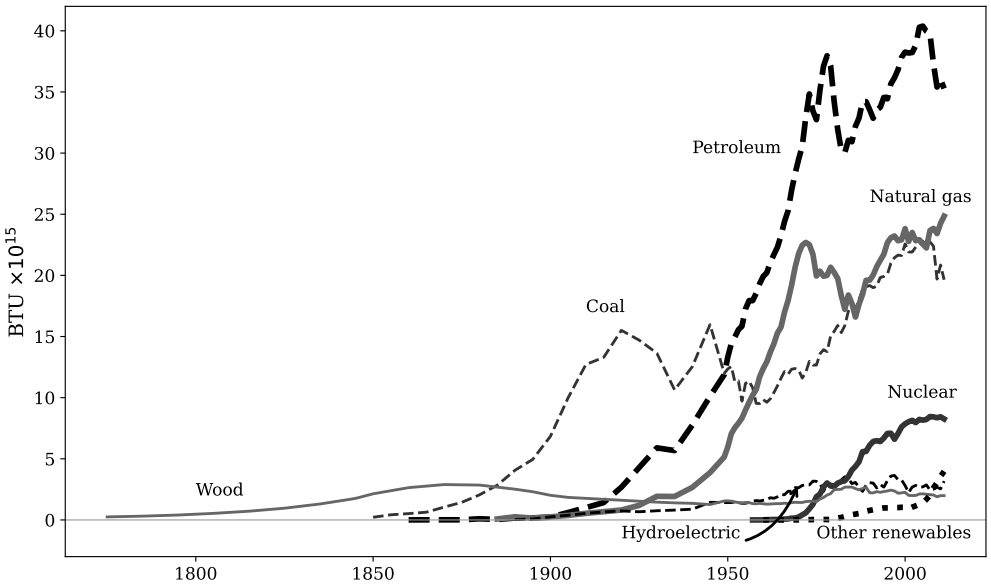


Figure 25.7: U.S. energy consumption.

Source:Administration (2012).

In per capita terms, the picture is different. For example, oil use peaked in the 1970s both in the U.S. and in the world as a whole and oil usage per capita has thus been decreasing since then. [23]

Turning to prices, we first look at a long time series of the oil price as displayed in Figure 25.8. Maybe the most prominent feature of the time series of the price is the enormous volatility, a feature that oil shares with other natural resources. Over a long horizon, the price appear to have no trend. However, if we limit attention to the postwar period, there is potentially a positive growth rate of the oil price.

Figure 25.9 shows the prices of coal, lead, zinc, and copper. Coal's price resembles that of oil, but the metals are somewhat different. They share the stationarity of a longer time horizon but they differ from the fossil fuels in that there is no visible upward trend in prices

[22]Hotelling, born in 1895 in Fulda, Minnesota, taught Milton Friedman statistics in the early 1930s. Friedman had a number of PhD students at Chicago, one of them Neil Wallace, the PhD advisor of one of the authors of the present chapter.

[23]See Hassler, Krusell, and Olovsson (2022).

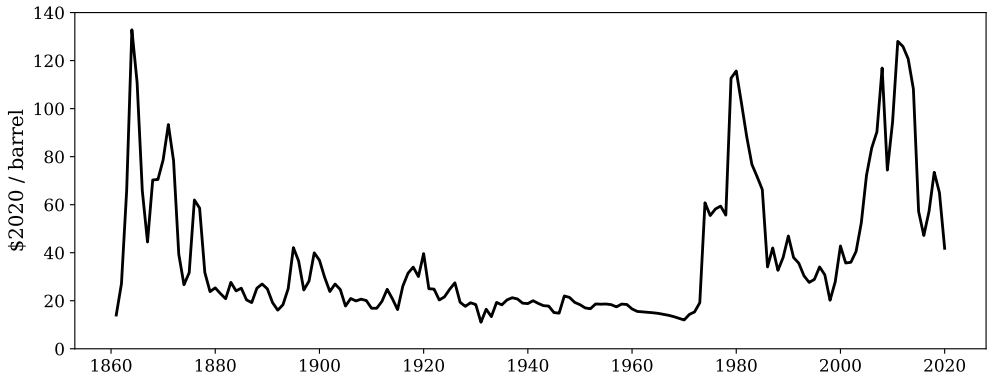


Figure 25.8: Crude oil price.

Source: BP (2021).

over the postwar period. A potential reason for this difference has to do with differences in extraction technologies across resources and over time. We will return to this point below.

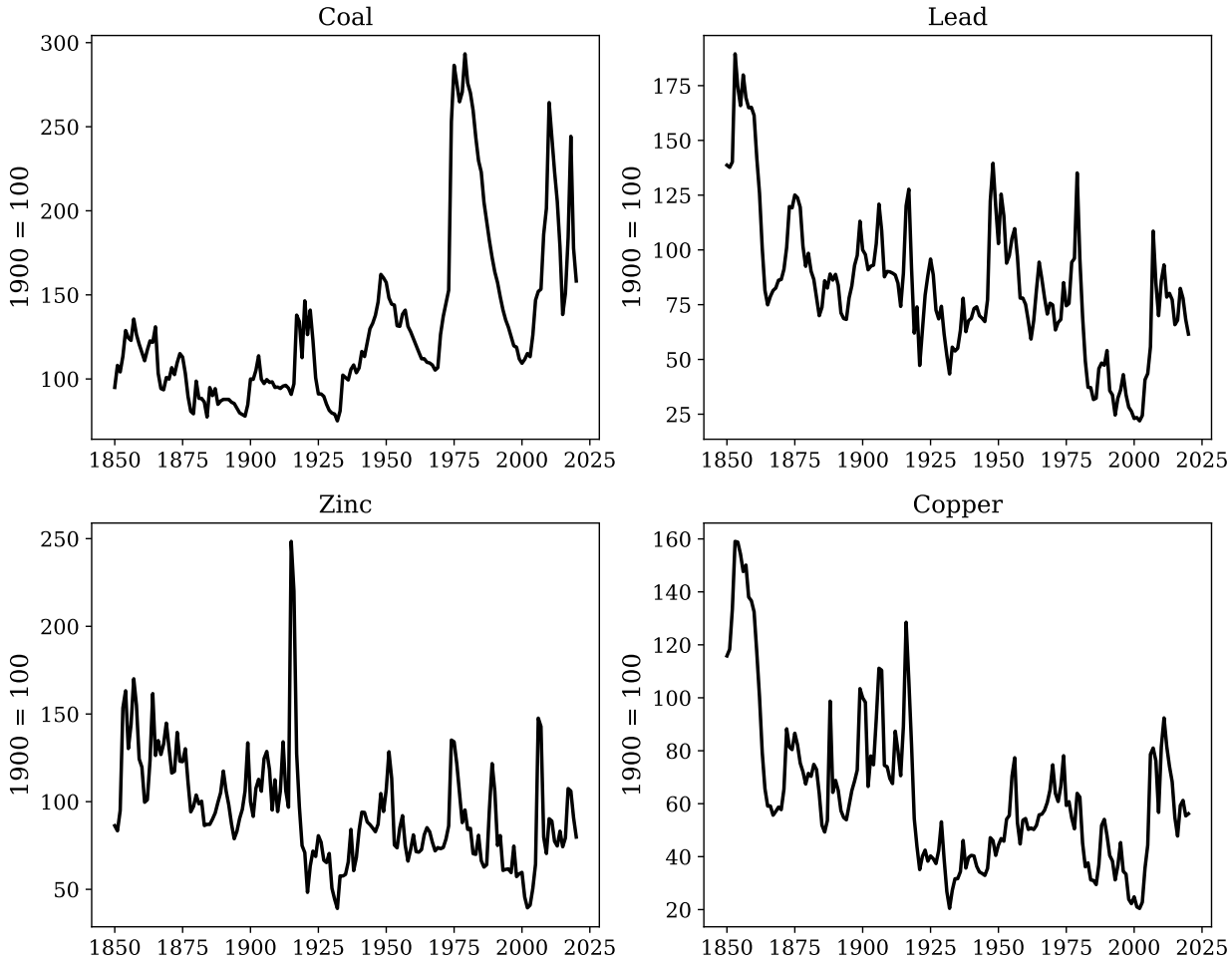


Figure 25.9: Inflation-adjusted prices for natural resources.

Source: Jacks (2019).

To summarize, we observe large swings in prices for natural resources, in fact on the order of magnitude of stock-price fluctuations. We do not observe marked trends, although for oil there is an upward trend over the postwar period. Quantities are rising significantly over time, fossil fuels included.[24] For fossil fuels, however, here we do see slowdowns, at least in per-capita consumption. It will turn out to be a challenge to account for these facts quantitatively.

25.5.2 Basic theory

The so called "cake-eating" problem that analyses how a finite resource should be depleted over time was first studied by Hotelling (1931) and later by Gale (1967). A more recent review of the empirical performance of Hotelling's theory can be found in Livernois (2009). The implications of natural resource scarcity then received renewed interest after the first oil-price shock in 1973 when concerns that world oil supply would run out or simply become too costly to use. The *Review of Economic Studies* featured a special issue on the economics of exhaustible resources already in 1974. The issue contains important contributions by Dasgupta and Heal, whose model we will use below, Solow, Stiglitz, among others; these papers are all concerned with the question of how economic growth is affected by the presence of an input that is depleted over time. Since a theory of endogenous technical change had not been developed at this point, however, the important concept of endogenous technical change, discussed in Chapter 13 and revisited in the present chapter, is naturally absent from the analysis in these papers.[25]

Pure cake eating: the planner Consider first the simplest possible cake eating problem: the cake's size is R_0, it can be stored over time and does not depreciate, and an infinitely-lived consumer with standard preferences likes cake. There is no other good in the economy. The planning problem thus reads

Consider a planning problem under zero extraction costs.

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log c_t$$

subject to

$$\sum_{t=0}^{\infty} c_t = R_0,$$

with c denoting consumption of the resource. Notice that there is no cost of production: this is the sense in which this is a cake. We assume logarithmic consumption mostly for illustrative purposes; we do, however, want to insist on balanced-growth preferences since this chapter concerns the long run. Thus, we could generalize preference somewhat, but not much.

[24] We omit data on quantities for lead, zinc, and copper; they display significant positive trends, too.

[25] Later contributions that are also concerned with the growth implications of natural resource scarcity but allow for endogenous technical change include Barbier (1999), Scholz and Ziemes (1999), Smulders and de Nooij (2003), Grimaud and Rouge (2003), and Hassler, Krusell, and Olovsson (2021a).

The solution to the planning problem is $c_t = (1-\beta)R_t$, where $R_{t+1} = R_t - c_t$, so that $c_t = (1-\beta)\beta^t R_0$.[26] The model is extremely stylized but allows us to make some important points. First, the problem is well defined, despite the resource use going to zero and the marginal utility of consumption going to infinity: utility is bounded. Second, consumption is higher in the beginning than later, all because of discounting.

Pure cake eating: the market outcome and Hotelling rents A third point we can learn by studying this simple economy comes from studying the market allocation. Let us imagine a perfectly competitive world where a representative consumer owns (and sells) the resource. Thus, the consumer would sell an amount of the resource e_t at time t at market price $\hat{p}^o_t$ ("o" for oil) in terms of the numéraire, which we take to be the resource at time 0, and also and buy an amount c_t at t, subject to the Arrow-Debreu budget constraint

$$\sum_{t=0}^{\infty} \hat{p}^o_t c_t = \sum_{t=0}^{\infty} \hat{p}^o_t e_t,$$

with $R_t - R_{t+1} = e_t \geq 0$ for all t. Market clearing in the simple cake-eating economy is then simply $c_t = e_t$ for all t.

Taking and combining first-order conditions with respect to c_t at two adjacent points in time, we obtain

$$\frac{\hat{p}^o_{t+1}}{\hat{p}^o_t} = \beta \frac{c_t}{c_{t+1}}.$$

The first-order conditions with respect to e_t and e_{t+1}, when combined, become $\hat{p}^o_t = \hat{p}^o_{t+1}$ for all t: it is equally costly to sell the resource at different points in time—the marginal production cost is zero—and therefore the price has to be the same; otherwise the consumer would sell all of the resource at the date with the highest price. Given this characterization of prices, we obtain, from the earlier first-order condition, that $c_{t+1} = \beta c_t$ or, imposing market clearing, $e_{t+1} = \beta e_t$. That is, markets deliver the same outcome as does the planning allocation, not surprisingly.

How should we interpret the price implications, i.e., that the price of the resource at different points in time is the same? Recall that these are Arrow-Debreu prices, in terms of the numéraire (cake at 0). If we instead think about spot prices, what would they be? There is no other consumption good than cake here, so the notion of a spot price is not a natural one. Imagine, however, that we had another consumption good whose time-t Arrow-Debreu price in terms of the numéraire would be p_t. Then p_t/p_{t+1} would be the gross real interest rate and optimal cake selling for the owner would require that the resource price satisfy the same condition as before, where p^o_t now means the spot price, in terms of the other consumption good, i.e., $\hat{p}^o_t = p^o_t p_t = p^o_{t+1} p_{t+1} = \hat{p}^o_{t+1}$, i.e.,

$$p^o_t = \frac{p_{t+1}}{p_t} p^o_{t+1}.$$

That is, the spot price of the resource today equals the *discounted* value of the spot price tomorrow. This is because consumption of the resource at all times requires indifference between selling the good at different points in time. By implication, the spot price of the resource would have to grow at the real rate of interest. This is the *Hotelling rule* governing the prices of resources in finite supply, for the case of zero extraction costs. This is a remarkably robust insight. The fact that the interest rate is zero in this simple economy is not central; if the interest rate were positive, the Hotelling rule would simply mean the price is growing.

[26] To obtain this amounts to a calculation that is standard for this text. Simply take first-order conditions with respect to c_t and c_{t+1} to see that, when the multiplier on the constraint is eliminated, $c_{t+1}/c_t = \beta$. Then substitute all the c_ts in terms of c_0 into the constraint.

So from the market analysis, we obtain this fourth result. A fifth result, which has been implicit in the discussion so far, is that $p_t^o > 0$ for all t: even though the cost of producing the resource is zero, the resource commands a positive price. This price, called the *Hotelling rent*, is a case where price exceeds marginal cost, even though there is perfect competition. If you earn a resource in limited supply, you earn rents from it, to the extent it is valued by consumers or firms.

A production economy Moving away from cake eating and toward a setting that views the natural resource as a input into production, consider planning a problem in the context of a growth model where the natural resource is used as an input: this is the Dasgupta and Heal (1974) setting. We assume that production is Cobb-Douglas in capital and energy, with decreasing returns to scale (labor is assumed fixed and would command the remainder of returns) and $\delta = 1$, i.e., capital depreciates fully from period to period. The resource, however, is modeled as before.

$$\max_{\{c_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log c_t$$

subject to

$$c_t + k_{t+1} = A z_t k_t^{\alpha} e_t^{\nu}$$

and

$$\sum_{t=0}^{\infty} e_t = R.$$

Our assumption of a Cobb-Douglas production function in k and e can be motivated if the resource costs, as a fraction of output, have been stationary. We will display the data later but a summary is that, for energy (or fossil fuel), the cost share fluctuates sharply over time, though without a strong trend. Hence, for long-run analysis, Cobb-Douglas seems like a decent starting point. For reasons we will explain shortly, we will also go beyond a Cobb-Douglas function.

The solution is straightforward to derive, following our basic chapters. For resource extraction we obtain the same solution here as for the cake-eating problem: $e_t = (1-\beta)R_t$, where $R_{t+1} = R_t - e_t$. Hence $e_t = (1-\beta)\beta^t R_0$.

As part of the solution, we find $k_{t+1} = \alpha\beta A z_t k_t^{\alpha} e_t^{\nu}$ and that long-run gross capital growth g is constant if z grows at a constant rate: $g = \gamma_z g^{\alpha}\beta^{\nu} = (\gamma_z \beta^{\nu})^{\frac{1}{1-\alpha}}$, where $\gamma_z - 1$ is the growth rate of z. For large enough γ_z, we see that $g > 1$. I.e., technology growth, of the form assumed here, allows net positive production growth.

As for price implications, from the Euler equation, we obtain $g = \beta(1+r)$, so that

$$1 + r = (\gamma_z \beta^{\nu})^{\frac{1}{1-\alpha}} / \beta.$$

Here, the net real interest rate is positive if $g > \beta$, i.e., if γ_z is large enough. The implications for the resource price, p^o, is $p_t = \nu y_t / e_t$, which grows at $g/\beta = 1 + r$. I.e., it rises at the real rate of interest. This result holds true time period by time period, as before. What the growth rate is of course depends on the value for the real interest rate; in the pure cake-eating case, the net real interest rate was zero at all times. Here, we can choose model

parameters so that it matches observations in the data. That would mean $1 + r > 1$, for at least most of the postwar period. Hence the resource price should rise exponentially, based on the simple setting here where there are no costs from extracting the resource.

Hotelling's rule in partial equilibrium with extraction costs Taking interest rates as given, it is straightforward to extend the Hotelling formula to a case where there are resource extraction costs, i.e., costs that resource producers have to pay in terms of goods. The idea is the same as before: for extraction to occur in two consecutive time periods, we need to have indifference between producing today and producing later. I.e., we need to have

$$p_t^o - mc_t = \frac{p_{t+1}^o - mc_{t+1}}{1 + r_t},$$

where mc_t is the "marginal extraction cost" at time t. This marginal cost should be interpreted in a broad sense: it involves the input costs of extraction at t but also potential dynamic effects of extraction: if, for example, it is more and more expensive to extract the more you extract, the marginal cost of extracting a unit now must include the cost of making extraction harder later. The specific forms for the marginal cost that would materialize depend on the detailed specification of the extraction technology, but there will always be a formula of the kind stated here (with a broad notion of mc).

This implies

$$\frac{p_{t+1}}{p_t} = 1 + r_t + \frac{1}{p_t}\left(mc_{t+1} - (1 + r_t)mc_t\right).$$

That is, if the marginal cost is rising faster than the rate of interest, the price has to rise faster to compensate. If we had data on marginal costs, we could assess whether this formula holds in the data, given a rate of interest, at least on average. As already pointed out, however, marginal costs are rarely directly observable and here they also have dynamic components, to the extent extraction at a point in time affects the extraction costs at other dates. The "given interest rate" is also a challenge, since it is not clear which interest rate to use: a low one from a comparatively safe rate of return (say, on government bonds), or a higher one that reflects risk? Many observers claim that the Hotelling rule is violated also in the postwar data, with reference to the lack of a trend whose average real increase equals real interest rate on the higher side.

Another challenge is volatility: the Hotelling-based price formula derived solely from a production perspective means a sharp focus on supply. One can imagine that various hard-to-predict shocks on the demand side cause fluctuations, but the literature has very few full quantitative general-equilibrium treatment of the world supply and demand for natural resources. For the purposes of studying fossil fuel and climate change, which is a very long-run issue, abstracting from short-run volatility seems a reasonable way forward.

25.5.3 Capital-energy complementarity and technical change

We discussed the challenges involved in accounting for the price data on natural resources. Another challenge, implicit in our discussion of the simple models above, is the strong positive trend in the use of natural resources over time, including fossil fuel. We mentioned that if the marginal production costs fell fast over time, then an upward trend would be visible. There are two factors working in opposite directions: fossil fuel is getting more and more difficult to extract, while technical change has made any given type of extraction cheaper over time. Arguably, the latter force does not dominate, or at least not so much that it can account for the strong positive trend in fossil-fuel use since the Industrial Revolution.

Another possibility arises, however, if one considers an alternative production technology, whereby energy—at least in a short-run sense—is very highly complementary with capital and labor. We now look at this case.

Strong complementarity For simplicity, consider an extreme case, a Leontief technology, whereby

$$y_t = \min\left\{Ak_t^\alpha, e_t\right\}.$$

Note here that A is not Hicksian but merely capital-augmenting. Also, to the extent at least one of the inputs is endogenously chosen, and costly to use, a Leontief function implies that $Ak_t^\alpha e_t$ must hold at all points in time. At time 0, this means that e_0 must be given by Ak_0^α, which is exogenous, k is a state variable. However, at time 1 and on, k is endogenous and costly: consumption earlier on needs to be forsaken in order to obtain it. What is the optimal path for capital and the resource? We will argue, without fully solving the model, that the outcome, based on a plausible parameter specification, is one where e initially grows over time and then falls.

So suppose that time 0 is at the beginning of the Industrial Revolution, where presumably k_0 is near zero, and A, interpreted as capital-specific technology, is also low. However, the energy problem had been "solved": for the purposes at the time, energy was cheap and plentiful. That is, the initial endowment R_0 was high in relation to Ak_0^α. Then the optimal allocation would be to accumulate capital over time, i.e., $k_0 < k_1 < k_2 \ldots$ at least for some time, so that $e_t = Ak_t^\alpha$ is increasing over time. Unlike in the standard neoclassical growth model, however, the economy would not asymptotically move toward a steady state, because e is in finite supply (R is high but not infinity). So eventually, e will have to decline, and k along with it. Fully working out a solution to this problem is somewhat involved but not necessary for the purposes here.

A more general case The example just studied makes clear that complementarity can generate an upward-sloping path for resource use. What would, however, a more systematic quantitative approach be like? A convenient generalization of the simple case displayed is a CES formulation:

$$y_t \equiv F\left(A_t k_t^\alpha l_t^{1-\alpha}, A_{et} e_t\right) = \left[(1-\nu)\left(A_t k_t^\alpha l_t^{1-\alpha}\right)^{\frac{\varepsilon-1}{\varepsilon}} + \nu\left(A_{et} e_t\right)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}}.$$

Here, we have reintroduced (possibly time-varying) labor, making clear that overall production has constant returns to scale; we have also allowed two technology variables, one augmenting each input, that can also change over time. The parameter ε is the elasticity of substitution between the capital-labor composite and the resource. The Leontief formulation used above is a special case: $\varepsilon = 0$.

Solow's growth-accounting method can be used to measure TFP changes over time, i.e., technological progress affecting overall production, under relatively weak assumptions.[27] However, how can *factor-specific* technology movements, such as those in A, and A_e be

[27] Recall, from Chapter 2, that the key assumptions are that inputs, outputs, and prices can be measured, that the production function has constant returns to scale, that there is perfect competition, and that firms maximize profits.

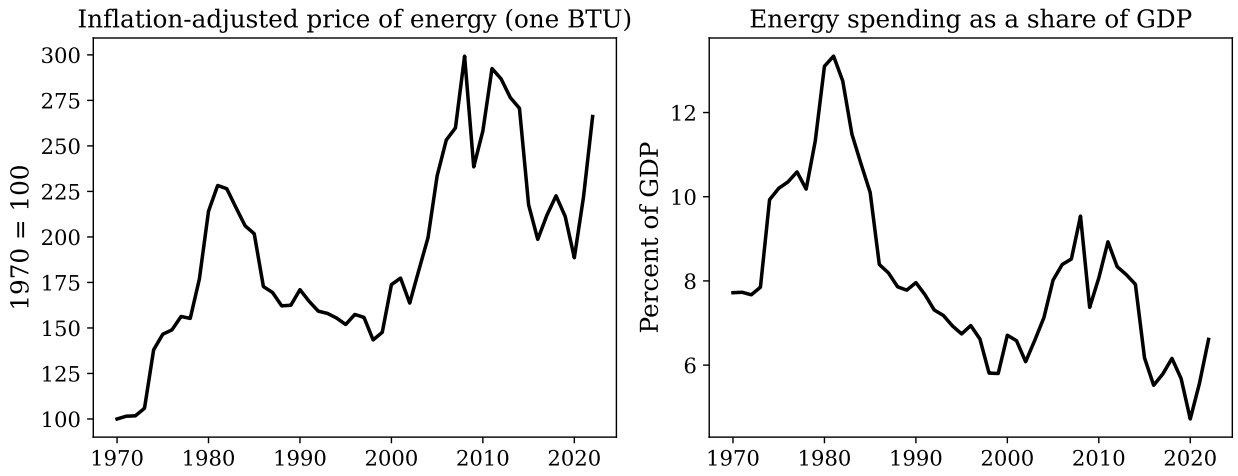


Figure 25.10: The price of fossil energy and its cost share.

measured? The short answer is that they can, under stronger assumptions on the production function. Suppose, for instance, that we knew that production was CES as just stated, and that we knew the parameters α, ν, and ε. Then the competitive firm's first-order conditions can be used to simply back out A_t and A_{et} period by period, given that all prices and quantities are observed. In this case, we obtain (after taking first-order conditions with respect to labor and the resource) and rearranging, that

$$A_t = \frac{y_t}{k_t^{\alpha} l_t^{1-\alpha}} \left(\frac{l_t^{share}}{1-\alpha} \right)^{\frac{\varepsilon}{\varepsilon-1}} \tag{25.33}$$

and

$$A_{et} = \frac{y_t}{e_t} \left(e_t^{share} \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{25.34}$$

where $l_t^{share} \equiv w_t l_t / y_t$ and $e_t^{share} = p_t^o e_t / y_t$ are cost shares.

An application to fossil fuel-based energy Hassler et al. (2021a) apply the ideas just discussed to the case of fossil fuel-based energy use in the United States. Relevant facts appear in Figure 25.10. Clearly, the price of fossil energy is extremely volatile over the medium run and the cost share of fossil energy appears to follow the price quite closely: when the price rises, so does the cost share. This means that a production function with ε close to zero fits the data very well; recall that a function with Cobb-Douglas or higher elasticity ($\varepsilon \geq 1$) would mean that when the price rises, the share falls (or, in the Cobb-Douglas case, does not move at all). In fact, even a Leontief function fits well. An α similar to what is used in the case without energy and a ν that matches average shares can then be added and, as a result, the movements of prices and shares match the data quite well. Given such a parameterization, it is then possible to use equations (25.33)–(25.34) to back out the factor-specific technologies. The result is depicted in Figure 25.11 together with the levels of y_t and e_t.

In Figure 25.11, e/l is (the normalized logarithm of) a fossil-fuel composite consisting of oil, coal, and natural gas per worker.[28]. Note that energy use per worker increased up to the

[28]The picture looks very similar if we instead plotted oil per person

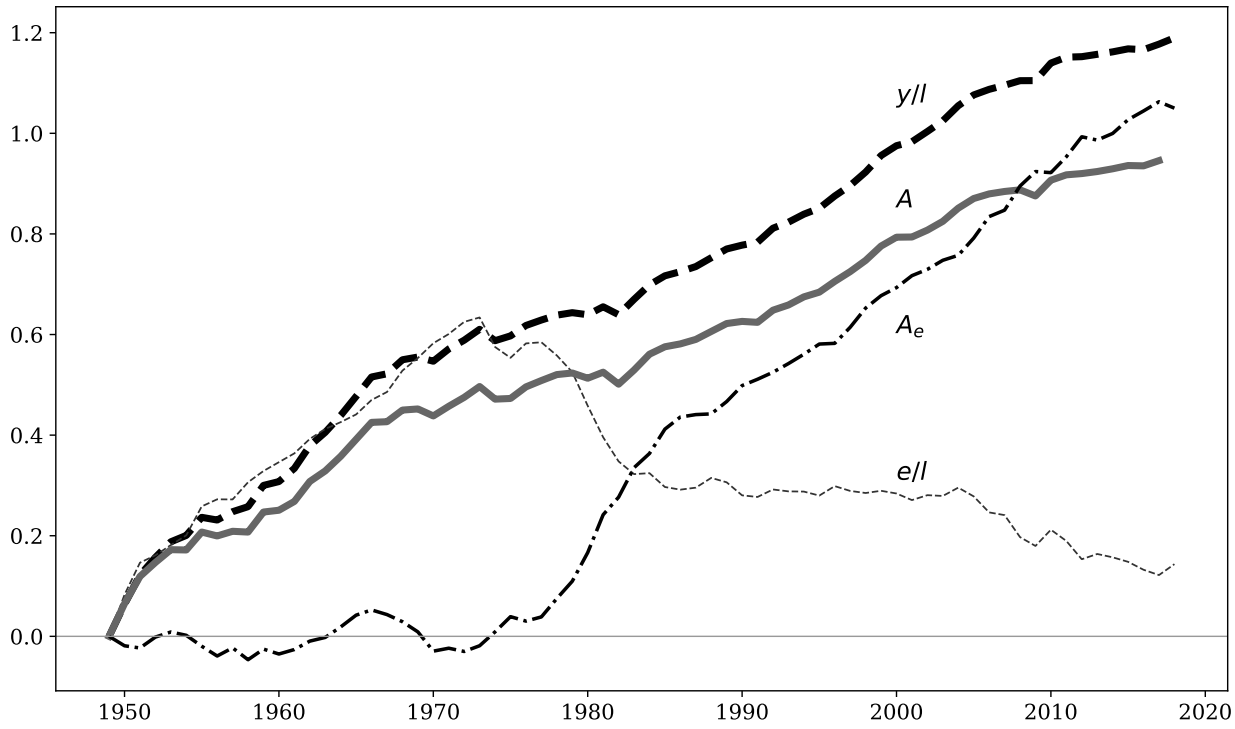


Figure 25.11: U.S. output, energy consumption and technological progress

early 1970s and it has been decreasing since. The variable y/l is (the normalized logarithm of) labor productivity. We see that e/l closely follows y/l between 1949 and 1973 and that, over this period, energy-saving technical change was effectively zero. Since 1973, however, A_e is increasing at a fast rate, while e/l has been falling. As a result, energy use in *efficiency units*, $A_e e/l$, has increased.[29] The reduction in physical energy use has thus been offset by faster energy-saving technical change.

The findings in Figure 25.11 strongly suggest that technology is endogenous: when prices of oil and gas shot up, energy-saving technological change started growing at a rapid rate. Moreover, technological change that augments capital and labor appears to have slowed down at the same time, suggesting that the endogeneity is *directed* in nature: away from the factors that are becoming relatively cheap toward those that are becoming relatively expensive.

Endogenous directed technology Chapter 13 developed models of endogenous technical change. An aspect that was not covered there is the multi-dimensionality of technology (such as the (A, A_e) here) and how inputs into research can be allocated based on relative profitability. In the context of Romer (1990) product variety model, A could thus be thought of as the variety range of intermediate inputs specifically used to save on capital and labor and whereas A_e would be the variety range of intermediate inputs directed toward energy-saving. Or, using Aghion and Howitt's (Aghion and Howitt, 1999) model of creative destruction, A and A_e could be the average efficiency levels of attained in (endogenous) productivity

[29] In fact, fossil-fuel use in efficiency units closely follows output.

enhancements of a fixed set of intermediate inputs that are capital/labor- and energy-saving, respectively. In each of these applications, one can imagine that there is a fixed set of R& D workers that can be allocated to either of the two innovative (and patent-seeking) sectors and that market forces would dictate where the R& D workers locate each period. A richer model would have the total amount of R& D workers be endogenous as well.

The technology menu approach In Hassler et al. (2021a), a simpler setting is used where a technology *menu* G is postulated that lists different options for each firm to choose freely among. In particular, consider

$$G\left(\frac{A_{t+1}}{A_t}, \frac{A_{e,t+1}}{A_{et}}\right) = 0, \tag{25.35}$$

where G is strictly increasing in both arguments. Then the idea is that a given, competitive firm can choose any $(A_{t+1}, A_{e,t+1})$ at time t satisfying equation (25.35) at no cost, treating $(A_t, A_{e,t})$ as given: these are the economy-wide average chosen in the previous period. I.e., all firms choose what respective technology saving they prefer next period but do not internalize the dynamic technology impact of their choice. Under some conditions on G (jointly with assumption on the firm's production function), the firm's profit maximization problem—which now involves choosing both inputs and the input-saving technology levels—is well defined and consistent with perfect competition. A simple, static example of this procedure is as follows: suppose (i) that the production function is CES in inputs x and y, i.e., $f(A_x x, A_y y)$, and that the elasticity of substitution between x and y is less than 1 (more complementary than Cobb-Douglas); and (ii) that $G(A_x, A_y)$ is a CES minus a positive constant, so that the non-linear (CES) average of A_x and A_y equals this constant. Then it is straightforward (though tedious) to show that the technology choice will lead to a "reduced-form" production function in x and y that is a CES in x and y with a higher elasticity of substitution than in f.[a] That is, when (A_x, A_y) is given and not subject to choice, x and y are harder to substitute than when they can be chosen. This captures the intuitive notion that the substitutability between inputs is higher in the long than in the short run. For example, fossil fuel may be very costly to replace as energy source in the short run but much less so in the longer run, since production techniques can then be developed that work better with other energy sources.

[a] The original paper working these results out is Leon-Ledesma and Satchi (2019).

Whichever way endogenous technology is added to the model, we can capture that natural notion that low short-run substitutability between energy and other input factors turns into higher substitutability in the longer run, when technology moves in such a way as to save on expensive inputs. Thus, a sharp short-run rise in the energy share in response to an increase in the price of energy will then be falling over time, even if the energy price remains high.

25.5.4 Taking stock

To conclude this section, we have presented some basic facts about natural resources in finite supply and some basic theory to begin accounting for the facts. We have focused on resources that, in contrast to the climate, have property rights associated to them, thus bracketing a broad class of environmental goods and services. The data we looked at showed

increasing paths of use, presumably going hand in hand with economic growth, but also a slowdown of use toward the end of the sample. This overall pattern, we argued, seems at least qualitatively consistent with the basic theory of resource depletion that we presented, so long as we incorporate time-varying, input-specific technology levels. The data also showed wildly fluctuating prices for natural resources, with weak time trends. The price data still remains a challenge to fully explain.

The basic theories we presented do not call for government intervention, at least not in an obvious way. The market outcome is optimal in the basic model, and although extensions involving endogenous directed technical change would suggest that some externalities to R&D be internalized with subsidies/taxes, it is not immediately obvious whether such subsidies should be directed or imply general to all kinds of R&D. Of course, the efficiency of the market outcome presumes that property owners and market participants are rational and that there are no major frictions in the relevant markets. The purpose of this section, however, was not to make the point that sustainability concerns generally are dealt with perfectly by markets; rather, it was to show how markets under ideal circumstances, given property rights, can be functioning well.

Bibliography

Abowd, J. and D. Card (1989). On the covariance structure of earnings and hours changes. *Econometrica 57*(2), 411–445.

Abramovitz, M. (1956). Resource and output trends in the united states since 1870. In *Resource and output trends in the United States since 1870*, pp. 1–23. NBER.

Acemoglu, D. (1998). Why do new technologies complement skills? directed technical change and wage inequality. *The quarterly journal of economics 113*(4), 1055–1089.

Acemoglu, D. (2008). *Introduction to modern economic growth*. Princeton university press.

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. Volume 4 of *Handbook of Labor Economics*, pp. 1043–1171. Elsevier.

Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American economic review 91*(5), 1369–1401.

Achdou, Y., J. Han, J.-M. Lasry, P.-L. Lions, and B. Moll (2022). Income and wealth distribution in macroeconomics: A continuous-time approach. *The Review of Economic Studies 89*(1), 45–86.

Adhami, M. (2025). Learning about spillovers from product dynamics. Unpublished manuscript, Stanford University.

Administration, T. E. I. (2012). The annual energy review 2011. Online document.

Aghion, P. and P. Howitt (1992). A model of growth through creative destruction. *Econometrica 60*(2), 323–351.

Aghion, P. and P. Howitt (1999). *Endogenous Growth Theory*. Cambridge, Massachusetts: MIT Press.

Aguiar, M. and M. Amador (2014). *Sovereign Debt*, pp. 647–87. North-Holland.

Aguiar, M. and M. Amador (2019). A contraction for sovereign debt models. *Journal of Economic Theory 183*, 842–875.

Aguiar, M. and M. Amador (2020). Self-fulfilling debt dilution: Maturity and multiplicity in debt models. *American Economic Review 110*(9), 2783–2818.

Aguiar, M., M. Amador, and S. Fourakis (2020). On the welfare losses from external sovereign borrowing. *IMF Economic Review 68*(4), 163–194. https://doi.org/10.1057/s41308-019-00103-2.

Aguiar, M., M. Amador, H. Hopenhayn, and I. Werning (2019). Take the short route: Equilibrium default and debt maturity. *Econometrica 87*(2), 423–462.

Aguiar, M., M. Bils, E. Hurst, and K. K. Charles (2021). Leisure luxuries and the labor supply of young men. *Journal of Political Economy 129*(2), 337–382.

Aguiar, M. and G. Gopinath (2006). Defaultable debt, interest rates and the current account. *Journal of International Economics 69*, 64–83.

Aguiar, M. and G. Gopinath (2007). Emerging market business cycles: The cycle is the trend. *Journal of Political Economy 115*, 69–102.

Aiyagari, S. R. (1994). Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics 109*(3), 659–684.

Aiyagari, S. R. and E. R. McGrattan (1998). The optimum quantity of debt. *Journal of Monetary Economics 42*(3), 447–469.

Akcigit, U. and S. T. Ates (2023). What happened to us business dynamism? *Journal of Political Economy 131*(8), 2059–2124.

Akcigit, U., S. T. Ates, and G. Impullitti (2018). Innovation and trade policy in a globalized world. Technical report, National Bureau of Economic Research.

Akitoby, B. and T. Stratmann (2008). Fiscal policy and financial markets. *The Economic Journal 118*, 1971–1985.

Albanesi, S. (2019). Changing business cycles: The role of women's employment. Technical report, National Bureau of Economic Research.

Alder, S., T. Boppart, and A. Müller (2022). A theory of structural change that can fit the data. *American Economic Journal: Macroeconomics 14*(2), 160–206.

Alfaro, L., A. Charlton, and F. Kanczuk (2009). Plant size distribution and cross-country income differences. In *NBER International seminar on macroeconomics*, Volume 5, pp. 243–272. The University of Chicago Press Chicago, IL.

Allais, M. (1947). *Economie et Interet*. Paris: Imprimerie Nationale.

Allen, F. (1985). Repeated principal-agent relationships with lending and borrowing. *Economics Letters 17*(1-2), 27–31.

Altinkilic, O. and R. S. Hansen (2000). Are there economies of scale in underwriting fees? evidence of rising external financial costs. *Review of Financial Studies 13*(1), 191–218.

Altonji, J. (1986a). Comment. *Econometric Reviews 5*(1), 147–151.

Altonji, J. G. (1986b). Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy 94*(3, Part 2), S176–S215.

Altug, S. and R. A. Miller (1990). Household choices in equilibrium. *Econometrica*, 543–570.

Alvarez-Parra, F., L. Brandao-Marques, and M. Toledo (2013). Durable goods, financial frictions, and business cycles in emerging economies. *Journal of Monetary Economics 60(6)*, 720–736.

Amador, M. and C. Phelan (2021). Reputation and sovereign default. *Econometrica 89*(4), 1979–2010.

An, S., Y. Chang, and S.-B. Kim (2009). Can a representative-agent model represent a heterogeneous-agent economy? *American Economic Journal: Macroeconomics 1*(2), 29–51.

Andolfatto, D. (1996). Business cycles and labor-market search. *American Economic Review 86*, 112–132.

Archer, D. (2005). Fate of fossil fuel CO_2 in geologic time. *Journal of Geophysical Research 110*(2).

Arellano, C. (2008). Default risk and income fluctuations in emerging economies. *American Economic Review 98(3)*, 690–712.

Arellano, C., Y. Bai, and P. Kehoe (2019). Financial frictions and fluctuations in volatility. *Journal of Political Economy 127*(5), 2049–2103.

Arellano, C., L. Bocola, and Y. Bai (2024). Sovereign default risk and firm heterogeneity. Working Paper.

Arellano, C. and A. Ramanarayanan (2012). Default and the maturity structure in sovereign bonds. *Journal of Political Economy 120*(2), 187–232.

Armington, P. S. (1969). A theory of demand for products distinguished by place of production. *IMF Staff Papers 16*(1), 159–178.

Aschauer, D. A. (1989). Is public expenditure productive? *Journal of Monetary Economics 23*(2), 177–200.

Asonuma, T., M. Chamon, and A. Sasahara (2016). Trade costs of sovereign debt restructurings: Does a market-friendly approach improve the outcome? IMF Working Paper No. 16/222.

Asonuma, T., D. Niepelt, and R. Ranciere (2017). Sovereign bond prices, haircuts and maturity. IMF Working Paper No. 17/119.

Asonuma, T. and C. Trebesch (2016). Sovereign debt restructurings: Preemptive or post-default? *Journal of the European Economic Association 14(1)*, 175–214.

Ates, S. T. and F. Saffie (2021). Fewer but better: Sudden stops, firm entry, and financial selection. *American Economic Journal: Macroeconomics 13*, 304–356.

Atkeson, A. and A. Burstein (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review 98*, 1998–2031.

Atkeson, A. and A. Burstein (2019). Aggregate implications of innovation policy. *Journal of Political Economy 127*(6), 2625–2683.

Atkeson, A., V. Chari, and P. Kehoe (1999). Taxing capital income: A bad idea. Technical Report 2331, Federal Reserve Bank of Minneapolis.

Attanasio, O. and S. J. Davis (1996). Relative wage movements and the distribution of consumption. *Journal of Political Economy 104*(6), 1227–1262.

Attanasio, O., H. Low, and V. Sánchez-Marcos (2008). Explaining changes in female labor supply in a life-cycle model. *American Economic Review 98*(4), 1517–1552.

Attanasio, O. and G. Weber (1993). Consumption growth, the interest rate and aggregation. *res 60*(3), 631–649.

Attanasio, O. and G. Weber (1995). Is consumption growth consistent with intertemporal optimisation? evidence from the consumption expenditure survey. *jpe 103*(6), 1121–1157.

Attanasio, O. P. (1999). Consumption. *Handbook of Macroeconomics 1*, 741–812.

Attanasio, O. P. and N. Pavoni (2011). Risk sharing in private information models with asset accumulation: Explaining the excess smoothness of consumption. *Econometrica 79*(4), 1027–1068.

Attanasio, O. P. and G. Weber (2010). Consumption and saving: models of intertemporal allocation and their implications for public policy. *Journal of Economic Literature 48*(3), 693–751.

Auclert, A., B. Bardóczy, M. Rognlie, and L. Straub (2021). Using the sequence-space jacobian to solve and estimate heterogeneous-agent models. *Econometrica 89*(5), 2375–2408.

Auclert, A. and K. Mitman (2018). Consumer bankruptcy as aggregate demand management. In *Society for Economic Dynamics Meeting Papers*.

Auclert, A. and M. Rognlie (2016). Unique equilibrium in the eaton-gersovitz model of sovereign debt. *Journal of Monetary Economics 84*, 134–146.

Bacchetta, P. and E. Van Wincoop (2006, June). Can information heterogeneity explain the exchange rate determination puzzle? *American Economic Review 96*(3), 552–576.

Bach, L., L. E. Calvet, and P. Sodini (2020, September). Rich pickings? risk, return, and skill in household wealth. *American Economic Review 110*(9), 2703–47.

Backus, D. K., P. J. Kehoe, and F. E. Kydland (1992, August). International real business cycles. *Journal of Political Economy 100*(4), 745–775.

Backus, D. K., P. J. Kehoe, and F. E. Kydland (1994a). Relative price movements in dynamic general equilibrium models of international trade. In F. Van der Ploeg (Ed.), *Handbook of International Macroeconomics*, Chapter 3, pp. 62–69. Basil Blackwell.

Backus, D. K., P. J. Kehoe, and F. E. Kydland (1994b). Dynamics of the trade balance and the terms of trade: The j-curve? *American Economic Review 84*(1), 84–103.

Bahaj, S. and R. Reis (2022, July). Central bank swap lines: Evidence on the effects of the lender of last resort. *The Review of Economic Studies 89*(4), 1654–1693.

Bai, Y., J.-V. Ríos-Rull, and K. Storesletten (2025). Demand shocks as productivity shocks. *Forthocoming, restud*.

Balasko, Y. and K. Shell (1980). The overlapping-generations model, i: The case of pure exchange without money. *Journal of Economic Theory 23*(3), 281–306.

Balke, N. and R. Gordon (1989). The estimation of prewar gross national product: Methodology and new evidence. *Journal of Political Economy 97*(1), 38–92.

Baqaee, D. R. and E. Farhi (2019). The macroeconomic impact of microeconomic shocks: Beyond hulten's theorem. *Econometrica 87*, 1155–1203.

Baqaee, D. R. and E. Farhi (2020). Productivity and misallocation in general equilibrium. *Quarterly Journal of Economics 135*, 105–163.

Barbier, E. (1999). Endogenous growth and natural resource scarcity. *Environmental and Resource Economic 14*, 51–74.

Barnett, W. A. (1980). Economic monetary aggregates an application of index number and aggregation theory. *Journal of Econometrics 14*(1), 11–48.

Barnichon, R. and G. Mesters (2020). Identifying modern macro equations with old shocks. *The Quarterly Journal of Economics 135*(4), 2255–2298.

Barro, R. J. (1974). Are government bonds net wealth? *Journal of Political Economy 82*(6), 1095–1117.

Barro, R. J. (1979). On the determination of public debt. *Journal of Political Economy 82*(6), 1095–1117.

Barro, R. J. and X. Sala-i Martin (1992). Convergence. *Journal of political Economy 100*(2), 223–251.

Barro, R. J. and X. Sala-i Martin (1995). *Economic growth*. MIT Press.

Barro, R. J. and X. Sala-i Martin (2004). *Economic growth*. MIT Press.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2013). Cross-country differences in productivity: The role of allocation and selection. *American economic review 103*(1), 305–334.

Baumol, W. J. (1986). Productivity growth, convergence, and welfare: what the long-run data show. *The american economic review*, 1072–1085.

Baxter, M. and M. Crucini (1995). Business cycles and the asset structure of foreign trade. *International Economic Review 36*(4), 821–54.

Baxter, M. and R. G. King (1999). Measuring business cycles: Approximate band-pass filters for economic time series. *The Review of Economics and Statistics 81*(4), 575–593.

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal 75*(299), 493–517.

Begenau, J. and J. Salomao (2019). Firm financing over the business cycle. *Review of Financial Studies 32*(4), 1235–1274.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *jpe 75*(4), 352–365.

Benhabib, J., A. Bisin, and M. Luo (2019). Wealth distribution and social mobility in the us: A quantitative approach. *American Economic Review 109*(5), 1623–1647.

Benigno, G. and P. Benigno (2003). Price stability in open economies. *The Review of Economic Studies 70*(4), 743–764.

Benigno, G. and C. Thoenissen (2008). Consumption and real exchange rates with incomplete markets and non-traded goods. *Journal of International Money and Finance 27*(6), 926–948.

Bergeaud, A., G. Cette, and R. Lecat (2016). Productivity trends in advanced countries between 1890 and 2012. *Review of Income and Wealth 62*(3), 420–444.

Bergin, P. R. and G. Corsetti (2020). Beyond competitive devaluations: The monetary dimensions of comparative advantage. *American Economic Journal: Macroeconomics 12*(4), 246–86.

Bernanke, B. S. (2015). The Taylor rule: A benchmark for monetary policy?

Bernanke, B. S. (2024). *Essays on the great depression*. Princeton University Press.

Bernanke, B. S. and A. S. Blinder (1992). The federal funds rate and the channels of monetary transmission. *The American Economic Review 82*(4), 901–921.

Bernanke, B. S. and M. Gertler (1989). Agency costs, net worth, and business fluctuations. *aer 79*(1), 14–31.

Bernanke, B. S., M. Gertler, and S. Gilchrist (1999). The financial accelerator in a quantitative business cycle framework. In J. B. Taylor and M. Woodford (Eds.), *Handbook of Macroeconomics*, Volume 1 of *Handbook of Macroeconomics*, Chapter 21, pp. 1341–1393. Elsevier.

Bewley, T. (1980). The optimum quantity of money. In J. Kareken and N. Wallace (Eds.), *Models of Monetary Economies*. Federal Reserve Bank of Minneapolis.

Bewley, T. (1983). A difficulty with the optimum quantity of money. *Econometrica*, 1485–1504.

BGR (2020). Bgr energy study 2019 – data and developments concerning german and global energy supplies (23). Online document.

Bianchi, M., B. R. Gudmundsson, and G. Zoega (2001). Iceland's natural experiment in supply-side economics. *American Economic Review 91*(5), 1564–1579.

Bick, A., N. Fuchs-Schündeln, and D. Lagakos (2018). How do hours worked vary with income? cross-country evidence and implications. *American Economic Review 108*(1), 170–199.

Bils, M., Y. Chang, and S.-B. Kim (2012). Comparative advantage and unemployment. *Journal of Monetary Economics 59*(2), 150–165.

Bils, M. and P. J. Klenow (2000). Does schooling cause growth? *American Economic Review 90*(5), 1160–1183.

Bils, M. and P. J. Klenow (2004a). Some evidence on the importance of sticky prices. *Journal of political economy 112*(5), 947–985.

Bils, M. and P. J. Klenow (2004b). Some evidence on the importance of sticky prices. *Journal of political economy 112*(5), 947–985.

Bippus, B., S. Lloyd, and D. Ostry (2023). Granular banking flows and exchange-rate dynamics. Bank of England working papers 1043, Bank of England.

Blanchard, J. O. and N. G. Mankiw (1988). Consumption: Beyond certainty equivalence. *American Economic Review 78*(2), 173–177.

Blanchard, O. (2023). *Fiscal Policy under Low Interest Rates*. The MIT Press.

Blanchard, O. and J. Galí (2007). Real wage rigidities and the new keynesian model. *Journal of Money, Credit and Banking 39*(s1), 35–65.

Blanchard, O., A. Leandro, and J. Zettelmeyer (2020). Redesigning the eu fiscal rules: From rules to standards. *Economic Policy*.

Blanchard, O. and R. Perotti (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *The Quarterly Journal of Economics 117*(4), 1329–1368.

Blanchard, O. J. (1987). Aggregate and individual price adjustment. *Brookings Papers on Economic Activity 18*(1987-1), 57–122.

Blanchard, O. J. and C. M. Kahn (1980). The solution of linear difference models under rational expectations. *Econometrica 48*(5), 1305–1311.

Blanchard, O. J. and D. Quah (1989). The dynamic effects of aggregate demand and supply disturbances. *American Economic Review 79*(4), 655–673.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Bloom, N., C. I. Jones, and J. Van Reenen (2020). Are ideas getting harder to find? *American Economic Review 110*(4), 1104–1144.

Bloom, N. and J. Van Reenen (2007). Measuring and explaining management practices across firms and countries. *The quarterly journal of Economics 122*(4), 1351–1408.

Blundell, R. and T. MaCurdy (1999). Labor supply: A review of alternative approaches. *Handbook of Labor Economics 3*, 1559–1695.

Blundell, R. and I. Preston (1998). Consumption inequality and income uncertainty. *The Quarterly Journal of Economics 113*(2), 603–640.

Bodenstein, M. (2011, July). Closing large open economy models. *Journal of International Economics 84*(2), 160–177.

Boehm, C. E., A. A. Levchenko, and N. Pandalai-Nayar (2023). The long and short (run) of trade elasticities. *American Economic Review 113*(4), 861–905.

Boivin, J., M. P. Giannoni, and I. Mihov (2009). Sticky prices and monetary policy: Evidence from disaggregated us data. *American economic review 99*(1), 350–84.

Bolton, P. and O. Jeanne (2009). Structuring and restructuring sovereign debt: The role of seniority. *Review of Economic Studies 76*, 879–902.

Boppart, T. (2014). Structural change and the kaldor facts in a growth model with relative price effects and non-gorman preferences. *Econometrica 82*(6), 2167–2196.

Boppart, T. and P. Krusell (2020). Labor supply in the past, present, and future: a balanced-growth perspective. *Journal of Political Economy 128*(1), 118–157.

Boppart, T., P. Krusell, and K. Mitman (2018). Exploiting MIT shocks in heterogeneous-agent economies: the impulse response as a numerical derivative. *Journal of Economic Dynamics and Control 89*(C), 68–92.

Boppart, T., P. Krusell, and J. Olsson (2024). Who should work how much? Working Paper 32977, National Bureau of Economic Research.

Bourguignon, F. and P.-A. Chiappori (1992). Collective models of household behavior: An introduction. In R. Blundell, I. Preston, and I. Walker (Eds.), *The Measurement of Household Welfare*, pp. 33–68. Cambridge University Press.

BP (2021). Statistical review of world energy 2021, 70th edition. Online document.

Brock, W. A. and L. J. Mirman (1972). Optimal economic growth and uncertainty: The discounted case. *Journal of Economic Theory 4*, 479–513.

Broda, C. and D. E. Weinstein (2006). Globalization and the gains from variety. *The Quarterly journal of economics 121*(2), 541–585.

Broda, C. and D. E. Weinstein (2010). Product creation and destruction: Evidence and price implications. *American Economic Review 100*(3), 691–723.

Broer, T. (2013). The wrong shape of insurance? what cross-sectional distributions tell us about models of consumption smoothing. *American Economic Journal: Macroeconomics 5*(4), 107–140.

Broer, T., N.-J. H. Hansen, P. Krusell, and E. Öberg (2020). The new keynesian transmission mechanism: A heterogeneous-agent perspective. *The Review of Economic Studies 87*, 77–101.

Broer, T., M. Kapička, and P. Klein (2017). Consumption risk sharing with private information and limited enforcement. *Review of Economic Dynamics 23*, 170–190.

Broner, F. A., A. Martin, and J. Ventura (2010). Sovereign risk and secondary markets. *American Economic Review 100(4)*, 1523–1555.

Browning, M. and T. F. Crossley (2001). The life-cycle model of consumption and saving. *Journal of Economic Perspectives 15*(3), 3–22.

Browning, M. and A. Lusardi (1996). Household saving: Micro theories and micro facts. *Journal of Economic Literature 34*(4), 1797–1855.

Brunnermeier, M. K. and Y. Sannikov (2014). A macroeconomic model with a financial sector. *American Economic Review 104*(2), 379–421.

Buera, F. J. and E. Oberfield (2020). The global diffusion of ideas. *Econometrica 88*(1), 83–114.

Bullard, J. and K. Mitra (2002). Learning about monetary policy rules. *Journal of monetary economics 49*(6), 1105–1129.

Bulow, J. and K. Rogoff (1989). Sovereign debt: Is to forgive to forget? *American Economic Review 79(1)*, 43–50.

Bureau, U. C. (1975). *Bicentennial Edition: Historical Statistics of the United States, Colonial Times to 1970*.

Caballero, R. J., E. Farhi, and P.-O. Gourinchas (2008). An equilibrium model of 'global imbalances' and low interest rates. *American Economic Review 98*(1), 358–393.

Calomiris, C. W. (1993). Financial factors in the great depression. *Journal of Economic Perspectives 7*(2), 61–85.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics 12*(3), 383–98.

Campbell, J. Y. (1987). Does saving anticipate declining labor income? an alternative test of the permanent income hypothesis. *Econometrica*, 1249–1273.

Campbell, J. Y. (2003). Consumption-based asset pricing. *Handbook of the Economics of Finance 1*, 803–887.

Campbell, J. Y. and N. G. Mankiw (1989). Consumption, income, and interest rates: Reinterpreting the time series evidence. *NBER Macroeconomics Annual 4*, 185–216.

Cantor, R. (1985). The consumption function and the precautionary demand for savings. *Economics Letters 17*(3), 207–210.

Cao, D., H. R. Hyatt, T. Mukoyama, and E. Sager (2022). Firm growth with new establishments. mimeo. Georgetown University, U.S. Census Bureau, and Federal Reserve Board.

Carleton, T., A. Jina, M. Delgado, M. Greenstone, T. Houser, S. Hsiang, A. Hultgren, R. E. Kopp, K. E. McCusker, I. Nath, J. Rising, A. Rode, H. K. Seo, A. Viaene, J. Yuan, and A. T. Zhang (2022). Valuing the global mortality consequences of climate change accounting for adaptation costs and benefits. *The Quarterly Journal of Economics 137*(4), 2037–2105.

Carlstrom, C. T. and T. S. Fuerst (1995). Interest rate rules vs. money growth rules. a welfare comparison in a cash-in-advance economy. *Journal of Monetary Economics 36*(2), 247–67.

Carroll, C. D. (1992). The buffer-stock theory of saving: Some macroeconomic evidence. *Brookings Papers on Economic Activity 1992*(2), 61–156.

Carroll, C. D. and M. S. Kimball (1996). On the concavity of the consumption function. *Econometrica*, 981–992.

Carvalho, V. M. and B. Grassi (2019). Large firm dynamics and the business cycle. *American Economic Review 109*, 1375–1425.

Cass, D. (1965). Optimum growth in an aggregative model of capital accumulation. *The Review of Economic Studies 32*(3), 233–240.

Castaneda, A., J. Diaz-Gimenez, and J.-V. Rios-Rull (2003). Accounting for the us earnings and wealth inequality. *Journal of Political Economy 111*(4), 818–857.

Castañeda, A., J. Díaz-Giménez, and J.-V. Ríos-Rull (2003). Accounting for the U.S. earnings and wealth inequality. *jpe 111*(4), 818–857.

Castañeda, A., J. Díaz-Giménez, and J.-V. Ríos-Rull (2003). Accounting for the U.S. earnings and wealth inequality. *The Journal of Political Economy 111*(4), 818–857.

Chahrour, R., V. Cormun, P. De Leo, P. A. Guerrón-Quintana, and R. Valchev (2024, June). Exchange rate disconnect revisited. Working Paper 32596, National Bureau of Economic Research.

Chamley, C. (1986). Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica 54*(3), 607–622.

Chang, Y. and S.-B. Kim (2006). From individual to aggregate labor supply: A quantitative analysis based on a heterogeneous agent macroeconomy. *International Economic Review 47*(1), 1–27.

Chang, Y. and S.-B. Kim (2007). Heterogeneity and aggregation: Implications for labor-market fluctuations. *American Economic Review 97*(5), 1939–1956.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007). Business cycle accounting. *Econometrica 75*(3), 781–836.

Chatterjee, S. (1994). Transitional dynamics and the distribution of wealth in a neoclassical growth model. *Journal of Public Economics 54*(1), 97–119.

Chatterjee, S., D. Corbae, K. Dempsey, and J. Ríos-Rull (2023). A quantitative theory of the credit score. *Econometrica 91*(5), 1803–1840.

Chatterjee, S., D. Corbae, and J.-V. Rios-Rull (2008). A finite-life private-information theory of unsecured consumer debt. *jet 142*(1), 149–177.

Chatterjee, S. and B. Eyigungor (2012). Maturity, indebtedness and default risk. *American Economic Review 102(6)*, 2674–2699.

Chatterjee, S. and B. Eyigungor (2015). Debt dilution and seniority in a model of defaultable sovereign debt. *American Economic Review*. Forthcoming.

Cheng, G., J. Diaz-Cassou, and A. Erce (2016). From debt collection to relief provision: 60 years of official debt restructurings through the paris club. ESM Working paper 2016-20.

Chetty, R. (2012). Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica 80*(3), 969–1018.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy 113*(1), 1–45.

Clarida, R. H. (1987). Consumption, liquidity constraints and asset accumulation in the presence of random income fluctuations. *International Economic Review*, 339–351.

Clower, R. (1967). A reconsideration of the microfoundations of monetary theory. *Economic Inquiry 6*(1), 1–8.

Cochrane, J. H. (1991). A simple test of consumption insurance. *Journal of Political Economy 99*(5), 957–976.

Cole, H. and P. J. Kehoe (1998). Models of sovereign debt: Partial versus general reputations. *International Economic Review 39(1)*, 55–70.

Cole, H. L. and N. R. Kocherlakota (2001). Efficient allocations with hidden income and hidden storage. *The Review of Economic Studies 68*(3), 523–542.

Cole, H. L. and M. Obstfeld (1991). Commodity trade and international risk sharing: How much do financial markets matter? *Journal of Monetary Economics 28*(1), 3–24.

Comin, D. and B. Hobijn (2010). An exploration of technology diffusion. *American economic review 100*(5), 2031–2059.

Comin, D. and M. Mestieri (2014). Technology diffusion: Measurement, causes, and consequences. In *Handbook of economic growth*, Volume 2, pp. 565–622. Elsevier.

Comin, D. and M. Mestieri (2018). If technology has arrived everywhere, why has income diverged? *American Economic Journal: Macroeconomics 10*(3), 137–178.

Cooley, T. and V. Quadrini (2003). Common currencies vs. monetary independence. *Review of Economic Studies 70*(4), 785–806.

Cooley, T. F., R. Marimon, and V. Quadrini (2004). Aggregate consequences of limited contracts enforceability. *Journal of Political Economy 111*(4), 421–46.

Cooley, T. F. and E. C. Prescott (1995). Economic growth and business cycles. In T. F. Cooley (Ed.), *Frontiers of Business Cycle Research*. Princeton University Press.

Corsetti, G., L. D'Aguanno, A. Dogan, S. Lloyd, and R. Sajedi (2023). Global value chains and international risk sharing. CEPR Discussion Paper DP18558, CEPR Press.

Corsetti, G. and L. Dedola (2005, September). A macroeconomic model of international price discrimination. *Journal of International Economics 67*(1), 129–155.

Corsetti, G., L. Dedola, and S. Leduc (2008a). High exchange-rate volatility and low pass-through. *Journal of Monetary Economics 55*(6), 1113–1128.

Corsetti, G., L. Dedola, and S. Leduc (2008b). International risk sharing and the transmission of productivity shocks. *Review of Economic Studies 75*(2), 443–473.

Corsetti, G., L. Dedola, and S. Leduc (2010). Optimal monetary policy in open economies. In B. M. Friedman and M. Woodford (Eds.), *Handbook of Monetary Economics*, Volume 3 of *Handbook of Monetary Economics*, Chapter 16, pp. 861–933. Elsevier.

Corsetti, G., L. Dedola, and S. Leduc (2014, February). The international dimension of productivity and demand shocks in the u.s. economy. *Journal of the European Economic Association 12*(1), 153–176.

Corsetti, G., L. Dedola, and S. Leduc (2023). Exchange rate misalignment and external imbalances: What is the optimal monetary policy response? *Journal of International Economics 114*, 103771.

Corsetti, G., A. Lipinska, and G. Lombardo (2025). International risk sharing and wealth allocation with higher cumulants. *Review of Economic Studies, Conditionally accepted*.

Corsetti, G. and P. Pesenti (2001). Welfare and macroeconomic interdependence. *The Quarterly Journal of Economics 116*(2), 421–445.

Covas, F. and W. denHaan (2011). The cyclical behavior of debt and equity finance.

Cruces, J. J. and C. Trebesch (2013). Sovereign defaults: The price of haircuts. *American Economic Journal: Macroeconomics 5(3)*, 85–117.

Crucini, M. and J. Davis (2016). Distribution capital and the short- and long-run import demand elasticity. *Journal of International Economics 100*(C), 203–219.

Cruz, J.-L. and E. Rossi-Hansberg (2022, May). Local carbon policy. (30027).

Dasgupta, P. and G. Heal (1974). The optimal depletion of exhaustible resources. *Review of Economic Studies 41*, 3–28.

Davis, S. J., J. C. Haltiwanger, and S. Schuh (1998). Job creation and destruction. *MIT Press Books 1*.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics 135*(2), 561–644.

De Nardi, M. and G. Fella (2017). Saving and wealth inequality. *Review of Economic Dynamics 26*, 280–300.

Deaton, A. (1992a). *Understanding consumption*. Oxford University Press.

Deaton, A. (1992b). *Understanding consumption*. Oxford University Press.

Decker, R. A., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2016). Declining business dynamism: What we know and the way forward. *American Economic Review 106*(5), 203–207.

Dell, M., B. F. Jones, and B. A. Olken (2014). What do we learn from the weather? the new climate–economy literature. *Journal of Economic Literature 52*(3), 740–798.

Devereux, M. B. and C. Engel (2002). Exchange rate pass-through, exchange rate volatility, and exchange rate disconnect. *Journal of Monetary Economics 49*(5), 913–940.

Devereux, M. B. and C. Engel (2003). Monetary policy in the open economy revisited: Price setting and exchange-rate flexibility. *Review of Economic Studies 70*(4), 765–783.

Dhyne, E., L. J. Alvarez, H. Le Bihan, G. Veronese, D. Dias, J. Hoffmann, N. Jonker, P. Lunnemann, F. Rumler, and J. Vilmunen (2006). Price changes in the euro area and the united states: Some facts from individual consumer price data. *Journal of Economic Perspectives 20*(2), 171–192.

Diamond, P. (1965). National debt in a neoclassical growth model. *aer 55*, 1126–50.

Diamond, P. (1982). Aggregate demand management in search equilibrium. *Journal of Political Economy 90*(5), 881–94.

Diamond, P. A. (1971). A model of price adjustment. *Journal of Economic Theory 3*, 156–168.

Dias, D. A. and C. Richmond (2009). Duration of capital market exclusion: An empirical investigation. Working Paper, UCLA.

Dixit, A. and J. Stiglitz (1977). Monopolistic competition and optimum product diversity. *American Economic Review 67*(3), 297–308.

Doepke, M. and R. M. Townsend (2006). Dynamic mechanism design with hidden income and hidden actions. *Journal of Economic Theory 126*(1), 235–285.

Domeij, D. and M. Floden (2006). The labor-supply elasticity and borrowing constraints: Why estimates are biased. *Review of Economic Dynamics 9*(2), 242–262.

Domeij, D. and J. Heathcote (2004). On the distributional effects of reducing capital taxes. *International Economic Review 45*(2), 523–554.

Drozd, L. A. and J. Nosal (2012). Understanding international prices: Customers as capital. *American Economic Review 102*(1), 364–95.

Du, W., C. E. Pflueger, and J. Schreger (2020). Sovereign debt portfolios, bond risks, and the credibility of monetary policy. *Journal of Finance 75(6)*, 3097–3138.

Du, W. and J. Schreger (2016). Local currency sovereign risk. *Journal of Finance 71(3)*, 1027–1070.

Dvorkin, M. A., J. M. Sanchez, H. Sapriza, and E. Yurdagul (2021). Sovereign debt restructurings. *American Economic Journal-Macroeconomics 13*(2), 26–77.

Eaton, J. and M. Gersovitz (1981). Debt with potential repudiation: theoretical and empirical analysis. *Review of Economic Studies 48*, 289–309.

Eaton, J. and S. Kortum (1996). Trade in ideas patenting and productivity in the oecd. *Journal of international Economics 40*(3-4), 251–278.

Eaton, J. and S. Kortum (1999). International technology diffusion: Theory and measurement. *International Economic Review 40*(3), 537–570.

Eckstein, Z. and E. Nagypal (2004). The evolution of u.s. earnings inequality: 1961−2002. *qr 28*(2), 10–29.

Egorov, K. and D. Mukhin (2023). Optimal policy under dollar pricing. *American Economic Review 113*(7), 1783–1824.

Eichenbaum, M., N. Jaimovich, and S. Rebelo (2011). Reference prices, costs, and nominal rigidities. *American Economic Review 101*(1), 234–62.

Eichengreen, B. and R. Hausmann (1999). Exchange rates and financial fragility. pp. 329–368. Proceedings –Economic Policy Symposium – Jackson Hole, Federal Reserve Bank of Kansas City.

Elsby, M. W. and G. Solon (2019). How prevalent is downward rigidity in nominal wages? international evidence from payroll records and pay slips. *Journal of Economic Perspectives 33*, 185–201.

Engel, C. and J. Park (2022). Debauchery and original sin: The currency composition of sovereign debt. *Journal of the European Economic Association 20(3)*, 1095–1144.

Erceg, C. J., D. W. Henderson, and A. T. Levin (2000). Optimal monetary policy with staggered wage and price contracts. *Journal of monetary Economics 46*(2), 281–313.

Erosa, A., L. Fuster, and G. Kambourov (2016). Towards a micro-founded theory of aggregate labour supply. *The Review of Economic Studies 83*(3), 1001–1039.

Evenson, R. E. and D. Gollin (2003). Assessing the impact of the green revolution, 1960 to 2000. *science 300*(5620), 758–762.

Fallick, B. and C. A. Fleischman (2004). Employer-to-employer flows in the u.s. labor market: The complete picture of gross worker flows. FEDS Working Papers 2004-34.

Farhi, E. and M. Maggiori (2017, 08). A model of the international monetary system. *The Quarterly Journal of Economics 133*(1), 295–355.

Feenstra, R. C., P. Luck, M. Obstfeld, and K. N. Russ (2018, March). In search of the armington elasticity. *The Review of Economics and Statistics 100*(1), 135–150.

Fehr, E. and L. Goette (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review 97*(1), 298–317.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, 3rd ed.* New York, NY: John Wiley and Sons.

Fernald, J. G. (2012). A quarterly, utilization-adjusted series on total factor productivity. Working Paper Series 2012-19, Federal Reserve Bank of San Francisco.

Fisher, J. D. and D. S. Johnson (2006). Consumption mobility in the united states: Evidence from two panel data sets. *Topics in Economic Analysis & Policy 6*(1).

Flavin, M. A. (1981). The adjustment of consumption to changing expectations about future income. *Journal of Political Economy 89*(5), 974–1009.

Folini, D., F. Kubler, A. Malova, and S. Scheidegger (2021). The climate in climate economics. Working paper.

Foster, L., J. Haltiwanger, and C. J. Krizan (2001). Aggregate productivity growth: Lessons from microeconomic evidence. In C. R. Hulten, E. R. Dean, , and M. J. Harper (Eds.), *New Developments in Productivity Analysis*. NBER.

Friedman, M. (1957). *Theory of the consumption function*. Princeton University Press.

Friedman, M. (1963). Inflation: Causes and consequences. In *Inflation: Causes and Consequences*. Asia Publishing House. Reprinted in *Dollars and Deficits: Inflation, Monetary Policy and the Balance of Payments,* by Milton Friedman, pp. 21-46. Englewood Cliffs, New Jersey: Prentice-Hall, 1968. As posted on *The Collected Works of Milton Friedman*, compiled and edited by Robert Leeson and Charles G. Palm. Hoover Institution Library and Archives.

Friedman, M. (1968). The role of monetary policy. *The American Economic Review 58*(1).

Fujita, S. and G. Ramey (2009). The cyclicality of separation and job finding rates. *International Economic Review 50*, 415–430.

Fukui, M. and T. Mukoyama (2024). Efficiency in job-ladder models. Boston University and Georgetown University.

Fukui, M., E. Nakamura, and J. Steinsson (2023). Women, wealth effects, and slow recoveries. *American Economic Journal: Macroeconomics 15*(1), 269–313.

Furman, J. and L. Summers (2020). A reconsideration of fiscal policy in the era of low interest rates. Working Paper.

Gabaix, X. (2009, 05). Power Laws in Economics and Finance. *Annual Review of Economics 1*(1), 255–294.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*, 733–772.

Gabaix, X. and M. Maggiori (2015). International liquidity and exchange rate dynamics. *The Quarterly Journal of Economics 130*(3), 1369–1420.

Galí, J. and T. Monacelli (2005, 07). Monetary policy and exchange rate volatility in a small open economy. *The Review of Economic Studies 72*(3), 707–734.

Gale, D. (1967). On optimal development in a multi-sector economy. *The Review of Economic Studies 34*, 1–18.

Galí, J. and L. Gambetti (2020). Has the us wage phillips curve flattened? a semi-structural exploration. In G. Castex, J. Galí, and D. Saravia (Eds.), *Changing Inflation Dynamics, Evolving Monetary Policy*, Volume 27 of *Series on Central Banking, Analysis, and Economic Policies*. Central Bank of Chile.

Galí, J. and M. Gertler (1999). Inflation dynamics: A structural econometric analysis. *Journal of monetary Economics 44*(2), 195–222.

Galí, J., J. D. López-Salido, and J. Vallés (2007). Understanding the effects of government spending on consumption. *Journal of the european economic association 5*(1), 227–270.

Gallup, J. L., J. D. Sachs, and A. D. Mellinger (1999). Geography and economic development. *International regional science review 22*(2), 179–232.

Galor, O. and D. Tsiddon (1997). The distribution of human capital and economic growth. *jeg 2*, 293–124.

García-Cicco, J., R. Pancrazi, and M. Uribe (2010). Real business cycles in emerging countries? *American Economic Review 100*, 2510–2531.

Garcia-Macia, D., C.-T. Hsieh, and P. J. Klenow (2019). How destructive is innovation? *Econometrica 87*(5), 1507–1541.

Ghironi, F. and M. J. Melitz (2005). International trade and macroeconomic dynamics with heterogeneous firms. *The Quarterly Journal of Economics 120*(3), 865–915.

Gilchrist, S., V. Yankov, and E. Zakrajsek (2009). Credit market shocks and economic fluctuations: Evidence from corporate bond and stock markets. *Journal of Monetary Economics 56*(4), 471–93.

Gollin, D., C. W. Hansen, and A. M. Wingender (2021). Two blades of grass: The impact of the green revolution. *Journal of Political Economy 129*(8), 2344–2384.

Golosov, M., M. Graber, M. Mogstad, and D. Novgorodsky (2021). How americans respond to idiosyncratic and exogenous changes in household wealth and unearned income. Technical report, National Bureau of Economic Research.

Golosov, M., J. Hassler, P. Krusell, and A. Tsyvinski (2014). Optimal taxes on fossil fuel in general equilibrium. *Econometrica 82*, 41–88.

Goodfriend, M. (1991). Interest rates and the conduct of monetary policy. *Carnegie-Rochester Conference Series on Public Policy 34*, 7–30.

Gopinath, G., E. Boz, C. Casas, F. J. Díez, P.-O. Gourinchas, and M. Plagborg-Møller (2020). Dominant Currency Paradigm. *American Economic Review 110*(3), 677–719.

Gordon, G. and S. Qiu (2018). A divide and conquer algorithm for exploiting policy function monotonicity. *Quantitative Economics 9*, 521–540.

Gordon, R. J. (2012, August). Is u.s. economic growth over? faltering innovation confronts the six headwinds. Working Paper 18315, National Bureau of Economic Research.

Gourinchas, P.-O. and H. Rey (2007). From world banker to world venture capitalist: Us external adjustment and the exorbitant privilege. In *G7 Current Account Imbalances: Sustainability and Adjustment*, NBER Chapters, pp. 11–66. National Bureau of Economic Research, Inc.

Greenwood, J., Z. Hercowitz, and G. W. Huffman (1988, June). Investment, capacity utilization, and the real business cycle. *American Economic Review 78*(3), 402–17.

Greenwood, J., Z. Hercowitz, and P. Krusell (1997). Long-run implications of investment-specific technological change. *The American economic review*, 342–362.

Greenwood, J., Z. Hercowitz, and P. Krusell (2000). The role of investment-specific technological change in the business cycle. *European Economic Review 44*((1)), 91–115.

Greenwood, J., A. Seshadri, and M. Yorukoglu (2005). Engines of liberation. *Review of Economic Studies 72*(1), 109–133.

Greenwood, J. and M. Yorokoglu (1997, April). 1974. *Carnegie-Rochester Conference Series on Public Policy 46*(2), 49–95.

Griffy, B. S. (2021, November). Search and the sources of life-cycle inequality. *International Economic Review 62*(4), 1321–1362.

Grigsby, J., E. Hurst, and A. Yildirmaz (2021). Aggregate nominal wage adjustments: New evidence from administrative payroll data. *American Economic Review 111*(2), 428–71.

Griliches, Z. (1969). Capital-skill complementarity. *reas 51*(4), 465–68.

Grimaud, A. and L. Rouge (2003). Non-renewable resources and growth with vertical innovations: optimum, equilibrium and economic policies. *Journal of Environmental Economics and Management 45*(5), 433–453.

Gronau, R. (1977). Leisure, home production, and work–the theory of the allocation of time revisited. *Journal of Political Economy 85*(6), 1099–1123.

Guerrieri, V. and G. Lorenzoni (2010). Credit crises, precautionary savings and the liquidity trap. Unpublished manuscript, University of Chicago Booth and Massacussetts Institute of Technology.

Guerrieri, V., G. Lorenzoni, L. Straub, and I. Werning (2022, May). Macroeconomic implications of covid-19: Can negative supply shocks cause demand shortages? *American Economic Review 112*(5), 1437–74.

Gürkaynak, R., B. Sack, and E. Swanson (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking 1*(1).

Hagedorn, M. and I. Manovskii (2008). The cyclical behavior of equilibrium unemployment and vacancies revisited. *American Economic Review 98*, 1692–1706.

Hall, G. J. and T. J. Sargent (2020, May). Debt and taxes in eight U.S. wars and two insurrections. Working Paper 27115, National Bureau of Economic Research.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of Political Economy 86*(6), 971–987.

Hall, R. E. (1988a, April). Intertemporal substitution in consumption. *Journal of Political Economy 96*(2), 339–57.

Hall, R. E. (1988b). Intertemporal substitution in consumption. *Journal of Political Economy 96*(2), 339–357.

Hall, R. E. (2005). Employment fluctuations with equilibrium wage stickiness. *American Economic Review 95*, 50–65.

Hall, R. E. and C. I. Jones (1999, 04). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics 114*(1), 83–116.

Hall, R. E. and D. W. Jorgenson (1969, June). Tax policy and investment behavior: Reply and further results. *American Economic Review 59*(3), 388–401.

Hamilton, J. D. (1994). *Time series analysis*. Princeton university press.

Hamilton, J. D. (2018). Why you should never use the hodrick-prescott filter. *Review of Economics and Statistics 100*(5), 831–843.

Hansen, G. D. (1985). Indivisible labor and the business cycle. *Journal of Monetary Economics 16*(3), 309–327.

Hansen, L. P. and R. J. Hodrick (1980, October). Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy 88*(5), 829–53.

Hansen, R. S. and P. Torregrosa (1992). Underwriter compensation and corporate monitoring. *Journal of Finance 47*(4), 1537–1555.

Hassler, J., P. Krusell, and C. Olovsson (2021a). Directed technical change as a response to natural resource scarcity. *Journal of Political Economy 129*(11), 3039–3072.

Hassler, J., P. Krusell, and C. Olovsson (2021b, 10). Presidential Address 2020 Suboptimal Climate Policy. *Journal of the European Economic Association 19*(6), 2895–2928.

Hassler, J., P. Krusell, and C. Olovsson (2022). Finite resources and the world economy. *Journal of International Economics 136*, 103592. NBER International Seminar on Macroeconomics 2021.

Hassler, J., P. Krusell, and C. Olovsson (2023). Climate policy in the wide world. Working paper.

Hatchondo, J. C. and L. Martinez (2009). Long-duration bonds and sovereign defaults. *Journal of International Economics 79*, 117–125.

Hatchondo, J. C., L. Martinez, and F. Roch (2022). Fiscal rules and the sovereign default premium. *American Economic Journal: Macroeconomics 14(4)*, 244–273.

Hatchondo, J. C., L. Martinez, and H. Sapriza (2010). Quantitative properties of sovereign default models: solution methods matter. *Review of Economic Dynamics 13*(4), 919–933.

Hatchondo, J. C., L. Martinez, and C. Sosa Padilla (2016). Debt dilution and sovereign default risk. *Journal of Political Economy 124(5)*, 1383–1422.

Hausmann, R. (2003). *Good credit ratios, bad credit ratings: The role of debt denomination*. London: Macmillan.

Hayashi, F. (1982). Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica: Journal of the Econometric Society*, 213–224.

Hayashi, F., J. Altonji, and L. Kotlikoff (1996). Risk-sharing between and within families. *Econometrica*, 261–294.

Heathcote, J. (2005). Fiscal policy with heterogenous agents and incomplete markets. *Review of Economic Studies 72*, 161–188.

Heathcote, J. and F. Perri (2002). Financial autarky and international business cycles. *Journal of Monetary Economics 49*(3), 601–627.

Heathcote, J., K. Storesletten, and G. L. Violante (2009). Quantitative macroeconomics with heterogeneous households. *Annual Review of Economics 1*(1), 319–354.

Heathcote, J., K. Storesletten, and G. L. Violante (2017). Optimal tax progressivity: An analytical framework. *Quarterly Journal of Economics 132*(4), 1693–1754.

Heer, B. and A. Maußner (2024). *Dynamic General Equilibrium Modeling, 3rd ed.* Cham, Switzerland: Springer.

Hennessy, C. A. and T. M. Whited (2007). How costly is external financing? evidence from a structural estimation. *Journal of Finance 62*(4), 1705–45.

Hopenhayn, H. (2014a). Firms, Misallocation, and Aggregate Productivity: A Review. *Annual Review of Economics 6*, 735–770.

Hopenhayn, H. (2014b). On the measure of distortions. NBER Working Paper 20404.

Hopenhayn, H. and R. Rogerson (1993). Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy 101*, 915–938.

Horn, S., C. M. Reinhart, and C. Trebesch (2021). China's overseas lending. *Journal of International Economics 133*, 103539.

Hosios, A. J. (1990). On the efficiency of matching and related models of search and unemployment. *Review of Economic Studies 57*, 279–298.

Hotelling, H. (1931). The economics of exhaustible resources. *Journal of Political Economy 39*(2), 137–175.

Hottman, C. J., S. J. Redding, and D. E. Weinstein (2016). Quantifying the sources of firm heterogeneity. *The Quarterly Journal of Economics 131*(3), 1291–1364.

Howard, P. H. and T. Sterner (2017). Few and not so far between: A meta-analysis of climate damage estimates. *Environmental and Resource Economics 68*, 197–225.

Howitt, P. (2000). Endogenous growth and cross-country income differences. *American Economic Review 90*(4), 829–846.

Hsieh, C.-T. and P. J. Klenow (2007). Relative prices and relative prosperity. *American Economic Review 97*(3), 562–585.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *Quarterly Journal of Economics 124*, 1403–1448.

Hsieh, C.-T. and P. J. Klenow (2014). The life cycle of plants in india and mexico. *The Quarterly Journal of Economics 129*(3), 1035–1084.

Hsieh, C.-T., P. J. Klenow, and I. Nath (2023). A global view of creative destruction. *Journal of Political Economy Macroeconomics 1*(2), 000–000.

Hsieh, C.-T., P. J. Klenow, and K. Shimizu (2022). Romer or ricardo? Technical report, Working paper.

Hsieh, C.-T. and E. Rossi-Hansberg (2023). The industrial revolution in services. *Journal of Political Economy Macroeconomics 1*(1), 3–42.

Hubmer, J., P. Krusell, and A. A. Smith (2018). A comprehensive quantitative theory of the u.s. wealth distribution. Mimeo, Yale University.

Huggett, M. (1993a). The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control 17*(5-6), 953–969.

Huggett, M. (1993b). The risk-free rate in heterogeneous-agent, incomplete-insurance economies. *jedc 17*(5), 953–969.

Huggett, M. (1997). The one-sector growth model with idiosyncratic shocks: Steady states and dynamics. *Journal of Monetary Economics 39*(3), 385–403.

Hulten, C. R. (1978). Growth Accounting with Intermediate Inputs. *Review of Economic Studies 45*, 511–518.

Huo, Z. and J.-V. Ríos-Rull (2015). The great recession and financial shocks. EPP, Federal Reserve Bank of Minneapolis.

Imai, S. and M. P. Keane (2004). Intertemporal labor supply and human capital accumulation. *International Economic Review 45*(2), 601–641.

IMF (2013). Staff guidance note on the application of the joint bank-fund debt sustainability framework for low-income countries. International Monetary Fund.

IMF (2022). Staff guidance note on the sovereign risk and debt sustainability framework for market access countries. International Monetary Fund.

IMF (2023). Coming down to earth: How to tackle soaring public debt. World Economic Outlook, Chapter 3, April.

Imrohoroğlu, A. (1989). Cost of business cycles with indivisibilities and liquidity constraints. *Journal of Political Economy 97*(6), 1364–1383.

Itskhoki, O. and D. Mukhin (2021a). Exchange rate disconnect in general equilibrium. *Journal of Political Economy 129*(8), 2183–2232.

Itskhoki, O. and D. Mukhin (2021b, June). Mussa puzzle redux. NBER Working Papers 28950, National Bureau of Economic Research, Inc.

Jacks, D. (2019). From boom to bust: A typology of real commodity prices in the long run. *Cliometrica 13*(2), 202–220.

Jaimovich, N. and S. Rebelo (2008, December). News and business cycles in open economies. *Journal of Money, Credit and Banking 40*(8), 1699–1711.

Jappelli, T. and L. Pistaferri (2006). Intertemporal choice and consumption mobility. *Journal of the European Economic Association 4*(1), 75–115.

Jappelli, T. and L. Pistaferri (2017). *The economics of consumption: theory and evidence*. Oxford University Press.

Jappelli, T. and L. Pistaferri (2020). Reported MPC and unobserved heterogeneity. *American Economic Journal: Economic Policy 12*(4), 275–297.

Jaramillo, L. and C. M. Tejada (2011). Sovereign credit ratings and spreads in emerging markets; does investment grade matter? *IMF Working Papers 11/44*.

Jeanne, O. and A. K. Rose (2002). Noise trading and exchange rate regimes. *The Quarterly Journal of Economics 117*(2), 537–569.

Jensen, M. K. (2018). Distributional comparative statics. *The Review of Economic Studies 85*(1), 581–610.

Jermann, U. and V. Quadrini (2012). Macroeconomic effects of financial shocks. *aer 102*(1), 238–71.

Johnson, D. S., J. A. Parker, and N. S. Souleles (2006, December). Household expenditure and the income tax rebates of 2001. *American Economic Review 96*(5), 1589–1610.

Johnson, R. C. (2014, October). Trade in intermediate inputs and business cycle comovement. *American Economic Journal: Macroeconomics 6*(4), 39–83.

Jones, C. I. (1995). R & d - based models of economic growth. *Journal of Political Economy 103*(4), 759–784.

Jones, C. I. and J. C. Williams (1998). Measuring the social return to r & d. *The Quarterly Journal of Economics 113*(4), 1119–1135.

Jones, C. I. and J. C. Williams (2000). Too much of a good thing? the economics of investment in r&d. *Journal of economic growth 5*, 65–85.

Jones, L. E. and R. Manuelli (1990). A convex model of equilibrium growth: Theory and policy implications. *Journal of political Economy 98*(5, Part 1), 1008–1038.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review 95*(1), 161–182.

Jordà, Ò., K. Knoll, D. Kuvshinov, M. Schularick, and A. M. Taylor (2019). The Rate of Return on Everything, 1870–2015. *The Quarterly Journal of Economics 134*(3), 1225–1298.

Jorgenson, D. W., M. S. Ho, and K. J. Stiroh (2008). A retrospective look at the us productivity growth resurgence. *Journal of Economic Perspectives 22*(1), 3–24.

Judd, K. L. (1985). Redistributive taxation in a simple perfect foresight model. *Journal of Public Economics 28*(1), 59–83.

Judd, K. L. (1998). *Numerical Methods in Economics*. MIT Press.

Kaas, L. (2021). Block-recursive equilibria in heterogeneous-agent models. mimeo. Goethe University Frankfurt.

Kahn, M. E., K. Mohaddes, R. N. C. Ng, M. H. Pesaran, M. Raissi, and J.-C. Yang (2019). Long-term macroeconomic effects of climate change: A cross-country analysis. Working paper.

Kahn, S. (1997). Evidence of nominal wage stickiness from microdata. *The American Economic Review 87*(5), 993–1008.

Kaldor, N. (1957). A model of economic growth. *The Economic Journal 67*(268), 591–624.

Kaplan, G. and G. Violante (2010). How much consumption insurance beyond self-insurance? *aejm 2*(4), 53–87.

Kaplan, G. and G. L. Violante (2014). A model of the consumption response to fiscal stimulus payments. *Econometrica 82*(4), 1199–1239.

Kaplan, G. and G. L. Violante (2018). Microeconomic heterogeneity and macroeconomic shocks. *Journal of Economic Perspectives 32*(3), 167–194.

Kaplan, G. and G. L. Violante (2022). The marginal propensity to consume in heterogeneous agent models. *Annual Review of Economics 14*, 747–775.

Kaplan, G., G. L. Violante, and J. Weidner (2014). The wealthy hand-to-mouth. *Brookings Papers on Economic Activity*, 121–154.

Karabarbounis, L. and B. Neiman (2014). The global decline of the labor share. *The Quarterly Journal of Economics 129*(1), 61–103.

Kareken, J. and N. Wallace (1981). On the indeterminacy of equilibrium exchange rates. *The Quarterly Journal of Economics 96*(2), 207–222.

Kartashova, K. (2014, October). Private equity premium puzzle revisited. *American Economic Review 104*(10), 3297–3334.

Katz, L. F. and K. M. Murphy (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *qje 107*(1), 35–78.

Keane, M. and R. Rogerson (2012). Micro and macro labor supply elasticities: A reassessment of conventional wisdom. *Journal of Economic Literature 50*(2), 464–476.

Keane, M. and R. Rogerson (2015). Reconciling micro and macro labor supply elasticities: A structural perspective. *Annu. Rev. Econ. 7*(1), 89–117.

Keane, M. P. (2011). Labor supply and taxes: A survey. *Journal of Economic Literature 49*(4), 961–1075.

Kehoe, T. J. and D. K. Levine (1993). Debt-constrained asset markets. *The Review of Economic Studies 60*(4), 865–888.

Kehoe, T. J. and D. K. Levine (2001). Liquidity constrained markets versus debt constrained markets. *Econometrica 69*(3), 575–598.

Kehoe, T. J., E. C. Prescott, et al. (2002). *Great depressions of the twentieth century*. Academic Press Cambridge, MA.

Keller, W. (2004). International technology diffusion. *Journal of economic literature 42*(3), 752–782.

Kendrick, J. W. (1961). *Productivity Trends in the United States*. Princeton University Press, Princeton, NJ.

Kesten, H. (1973). Random Difference Equations and Renewal Theory for Products of Random Matrices. *Acta Mathematica 131*(1), 207–248.

Keynes, J. M. (1936). *The General Theory of Interest, Employment and Money*. New York: Harcourt, Brace & World.

Khan, A. and J. Thomas (2011). Credit shocks and aggregate fluctuations in an economy with production heterogeneity. Unpublished Manuscript, Department of Economics, Ohio State University.

King, R. and S. Rebelo (1993). Transitional dynamics and economic growth in the neoclassical model. *American Economic Review 83*(4), 908–931.

King, R. G., C. I. Plosser, and S. T. Rebelo (1988). Production, growth and business cycles: I. the basic neoclassical model. *Journal of Monetary Economics 21*(2-3), 195–232.

Kiyotaki, N. and J. Moore (1997, April). Credit cycles. *Journal of Political Economy 105*(2), 211–48.

Kiyotaki, N. and R. Wright (1989). On money as a medium of exchange. *Journal of Political Economy 97*(4), 927–54.

Klein, P., P. Krusell, and J.-V. Ríos-Rull (2008, 07). Time-consistent public policy. *The Review of Economic Studies 75*(3), 789–808.

Klenow, P. J. and O. Kryvtsov (2008). State-dependent or time-dependent pricing: Does it matter for recent us inflation? *The Quarterly Journal of Economics 123*(3), 863–904.

Klenow, P. J. and H. Li (2021). Innovative growth accounting. *NBER Macroeconomics Annual 35*(1), 245–295.

Klenow, P. J. and B. A. Malin (2010). Microeconomic evidence on price-setting. In *Handbook of monetary economics*, Volume 3, pp. 231–284. Elsevier.

Klenow, P. J. and A. Rodriguez-Clare (1997). The neoclassical revival in growth economics: Has it gone too far? *NBER Macroeconomics Annual 12*, 73–103.

Klenow, P. J. and A. Rodriguez-Clare (2005). Externalities and growth. *Handbook of economic growth 1*, 817–861.

Klette, T. J. and S. Kortum (2004). Innovating firms and aggregate innovation. *Journal of Political Economy 112*, 986–1018.

Kocherlakota, N. R. (1996). Implications of efficient risk sharing without commitment. *The Review of Economic Studies 63*(4), 595–609.

Kongsamut, P., S. Rebelo, and D. Xie (2001). Beyond balanced growth. *The Review of Economic Studies 68*(4), 869–882.

Koopmans, T. C. (1963). On the concept of optimal economic growth. Cowles Foundation Discussion Papers 163, Cowles Foundation for Research in Economics, Yale University.

Kopecky, K. A. and R. M. Suen (2010). Finite state markov-chain approximations to highly persistent processes. *Review of Economic Dynamics 13*, 701–704.

Kremer, M. (1993). The o-ring theory of economic development. *The quarterly journal of economics 108*(3), 551–575.

Kremer, M., J. Willis, and Y. You (2022). Converging to convergence. *NBER macroeconomics annual 36*(1), 337–412.

Krueger, D. and H. Lustig (2010). When is market incompleteness irrelevant for the price of aggregate risk (and when is it not)? *Journal of Economic Theory 145*(1), 1–41.

Krueger, D., K. Mitman, and F. Perri (2016). Chapter 11 - macroeconomics and household heterogeneity. Volume 2 of *Handbook of Macroeconomics*, pp. 843–921. Elsevier.

Krueger, D. and F. Perri (2006). Does income inequality lead to consumption inequality? Evidence and theory. *The Review of Economic Studies 73*(1), 163–193.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *The American Economic Review 70*(5), 950–959.

Krusell, P., T. Mukoyama, and A. Şahin (2010). Labour-market matching with precautionary savings and aggregate fluctuations. *Review of Economic Studies 77*, 1477–1507.

Krusell, P., T. Mukoyama, R. Rogerson, and A. Şahin (2017). Gross worker flows over the business cycle. *American Economic Review 107*, 3447–3476.

Krusell, P., T. Mukoyama, A. Şahin, and A. A. Smith Jr (2009). Revisiting the welfare effects of eliminating business cycles. *Review of Economic Dynamics 12*(3), 393–404.

Krusell, P., L. E. Ohanian, J. V. Ríos-Rull, and G. L. Violante (2000). Capital-skill complementarity and inequality: A macroeconomic analysis. *Econometrica 68(5)*, 1029–1054.

Krusell, P. and A. A. Smith, Jr (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy 106*(5), 867–896.

Krusell, P. and A. A. J. Smith (2022). Climate change around the world. Working paper.

Krusell, P. and T. Smith (2014). Is piketty's second law of capitalism fundamental? Unpublished Manuscriipt, Institute for International Economic Studies.

Kuhn, M. and J.-V. Rios-Rull (2025). Income and wealth inequality in the united states: An update including the 2022 wave.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of monetary economics 47*(3), 523–544.

Kydland, F. and E. Prescott (2004). Contribution to dynamic macroeconomics: The time consistency of economic policy and the driving forces behind business cycles. *Royal Swedish Academy of Sciences*.

Kydland, F. E. and E. C. Prescott (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy 85*(3), 473–491.

Kydland, F. E. and E. C. Prescott (1982a). Time to build and aggregate fluctuations. *eco 50*(6), 1345–1370.

Kydland, F. E. and E. C. Prescott (1982b, November). Time to build and aggregate fluctuations. *Econometrica 50*(6), 1345–70.

Lagakos, D., B. Moll, T. Porzio, N. Qian, and T. Schoellman (2018). Life cycle wage growth across countries. *Journal of Political Economy 126*(2), 797–849.

Lagos, R. and R. Wright (2005). A unified framework for monetary theory and policy analysis. *Journal of Political Economy 113*(3), 463–484.

Lang, V., D. Mihalyi, and A. F. Presbitero (2023). Borrowing costs after sovereign debt relief. *American Economic Journal: Economic Policy 15*, 331–358.

LaSalle, J. P. (1986). *The Stability and Control of Discrete Processes*. New York: Springer-Verlag.

Lee, Y. and T. Mukoyama (2018). A model of entry, exit, and plant-level dynamics over the business cycle. *Journal of Economic Dynamics and Control 96*, 1–25.

Leeper, E. M., T. B. Walker, and S.-C. Yang (2010). Government investment and fiscal stimulus. *Journal of Monetary Economics 57*, 253–92.

Leland, H. E. (1968). Saving and uncertainty: The precautionary demand for saving. *The Quarterly Journal of Economics 82*(3), 465–473.

Leon-Ledesma, M. and M. Satchi (2019). Appropriate technology and balanced growth. *Review of Economic Studies 86*(2), 807–835.

Li, D., M. Plagborg-Møller, and C. K. Wolf (2024). Local projections vs. vars: Lessons from thousands of dgps. *Journal of Econometrics*, 105722.

Lintner, J. (1956). Distribution of incomes of corporations among dividends, retained earnings, and taxes. *46*(2), 97–113.

Livernois, J. (2009). On the empirical significance of the hotelling rule. *Review of Environmental Economics and Policy 3*(1), 22–41.

Livshits, I., J. MacGee, and M. Tertilt (2007). Consumer bankruptcy: A fresh start. *aer 97*(1), 402–418.

Ljungqvist, L. and T. J. Sargent (2006). Time-averaging and labor supply. *Journal of Monetary Economics 53*(2), 233–259.

Ljungqvist, L. and T. J. Sargent (2012). *Recursive Macroeconomic Theory, 3rd Ed.* Cambridge, Massachusetts: MIT Press.

Ljungqvist, L. and T. J. Sargent (2018). *Recursive macroeconomic theory*. MIT press.

Low, H. W. (2005). Self-insurance in a life-cycle model of labour supply and savings. *Review of Economic Dynamics 8*(4), 945–975.

Lucas, R. (1982). Interest rates and currency prices in a two-country world. *Journal of Monetary Economics 10*(3), 335–359.

Lucas, R. E. (1972). Expectations and the neutrality of money. *Journal of economic theory 4*(2), 103–124.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, Volume 1, pp. 19–46. North-Holland.

Lucas, R. E. (1977). Understanding business cycles. *Carnegie-Rochester Conference Series on Public Policy 5*, 7–29.

Lucas, R. E. (1980). Methods and problems in business cycle theory. *Journal of Money, Credit and banking 12*(4), 696–715.

Lucas, R. E. and J. P. Nicolini (2015). On the stability of money demand. *Journal of Monetary Economics 73*, 48–65. Carnegie-Rochester-NYU Conference Series on Public Policy "Monetary Policy: An Unprecedented Predicament" held at the Tepper School of Business, Carnegie Mellon University, November 14-15, 2014.

Lucas, R. E. and L. A. Rapping (1969). Real wages, employment, and inflation. *Journal of political economy 77*(5), 721–754.

Lucas, R. E. and N. L. Stokey (1983). Optimal fiscal and monetary policy in an economy without capital. *Journal of Monetary Economics 12*(1), 55–93.

Lucas Jr, R. E. (1988). On the mechanics of economic development. *Journal of monetary economics 22*(1), 3–42.

Luttmer, E. G. (2011). On the mechanics of firm growth. *Review of Economic Studies 78*, 1042–1068.

Mace, B. J. (1991). Full insurance in the presence of aggregate uncertainty. *Journal of Political Economy 99*(5), 928–956.

MaCurdy, T. E. (1981). An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy 89*(6), 1059–1085.

Mankiw, N. G. and R. Reis (2002, 11). Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve*. *The Quarterly Journal of Economics 117*(4), 1295–1328.

Mankiw, N. G., D. Romer, and D. N. Weil (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics 107*(2), 407–437.

Marcet, A. and M. O. Ravn (2004). The hp-filter in cross-country comparisons. *Available at SSRN 511369*.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. New York, New York: Oxford University Press.

Mateos-Planas, X., S. McCrary, J.-V. Ríos-Rull, and A. Wicht (2025). Commitment in the canonical sovereign default model. *Journal of International Economics 157*, 104120.

Mavroeidis, S., M. Plagborg-Møller, and J. H. Stock (2014, March). Empirical evidence on inflation expectations in the new keynesian phillips curve. *Journal of Economic Literature 52*(1), 124–88.

Mazzocco, M. and S. Saini (2012). Testing efficient risk sharing with heterogeneous risk preferences. *American Economic Review 102*(1), 428–468.

McCall, J. J. (1970). Economics of information and job search. *Quarterly Journal of Economics 84*, 113–126.

McLaughlin, K. J. (1994). Rigid wages? *Journal of Monetary Economics 34*, 383–414.

McLeay, M. and S. Tenreyro (2025). Dollar dominance and the transmission of monetary policy. *The Quarterly Journal of Economics*. Forthcoming.

Meese, R. and K. S. Rogoff (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics 14*, 345–373.

Meghir, C. and L. Pistaferri (2011). Earnings, consumption and life cycle choices. In *Handbook of Labor Economics*, Volume 4, pp. 773–854. Elsevier.

Mehra, R. and E. C. Prescott (1985). The equity premium: A puzzle. *Journal of Monetary Economics 15*(2), 145–161.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *econometrica 71*(6), 1695–1725.

Mendelsohn, R., W. Nordhaus, and D. Shaw (1994). The impact of global warming on agriculture: A ricardian analysis. *American Economic Review 84*(4), 753–771.

Mendoza, E. and V. Yue (2012). A general equilibrium model of sovereign default and business cycles. *The Quarterly Journal of Economics 127(2)*, 889–946.

Merz, M. (1995). Search in the labor market and the real business cycle. *Journal of Monetary Economics 36*, 269–300.

Miller, B. L. (1974). Optimal consumption with a stochastic income stream. *Econometrica*, 253–266.

Miller, N. H. (2025). Industrial organization and the rise of market power. *International Journal of Industrial Organization 103131*.

Mincer, J. A. (1974). The human capital earnings function. In *Schooling, experience, and earnings*, pp. 83–96. NBER.

Miranda, M. J. and P. L. Fackler (2002). *Applied Computational Economics and Finance*. MIT Press.

Miranda-Agrippino, S. and H. Rey (2020). U.s. monetary policy and the global financial cycle. *Review of Economic Studies 87*(6), 2754–2776.

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *res 38*(2), 175–208.

Mitman, K. and S. Rabinovich (2019). Do unemployment benefit extensions explain the emergence of jobless recoveries?

Modigliani, F. and R. Brumberg (1954). Utility analysis and the consumption function: An interpretation of cross-section data. *Franco Modigliani 1*(1), 388–436.

Modigliani, F. and M. H. Miller (1958). The cost of capital, corporate finance and the theory of investment. *aer 48*(3), 261–279.

Moscoso Boedo, H. J. and T. Mukoyama (2012). Evaluating the effects of entry regulations and firing costs on international income differences. *Journal of Economic Growth 17*, 143–170.

Mountford, A. and H. Uhlig (2009). What are the effects of fiscal policy shocks? *Journal of applied econometrics 24*(6), 960–992.

Muellbauer, J. (1994). The assessment: consumer expenditure. *Oxford Review of Economic Policy 10*(2), 1–41.

Mukoyama, T. and S. Osotimehin (2019). Barriers to reallocation and economic growth: The effects of firing costs. *American Economic Journal: Macroeconomics 11*, 235–270.

Mukoyama, T. and L. Popov (2014). The Political Economy of Entry Barriers. *Review of Economic Dynamics 17*, 383–416.

Mussa, M. (1986, January). Nominal exchange rate regimes and the behavior of real exchange rates: Evidence and implications. *Carnegie-Rochester Conference Series on Public Policy 25*(1), 117–214.

Nakamura, E. and J. Steinsson (2008). Five facts about prices: A reevaluation of menu cost models. *The Quarterly Journal of Economics 123*(4), 1415–1464.

Nakamura, E. and J. Steinsson (2014). Fiscal stimulus in a monetary union: Evidence from us regions. *American Economic Review 104*(3), 753–792.

Nam, D. and J. Wang (2015). The effects of surprise and anticipated technology changes on international relative prices and trade. *Journal of International Economics 97*(1), 162–177.

Nath, I. B., V. A. Ramey, and P. J. Klenow (2023). How much will global warming cool global growth? Technical report, Working paper.

Nelson, J. A. (1994). On testing for full insurance using consumer expenditure survey data. *Journal of Political Economy 102*(2), 384–394.

Neumeyer, P. and F. Perri (2005). Business cycles in emerging economies: the role of interest rates. *Journal of Monetary Economics 52*, 345–380.

Ngai, L. R. and C. A. Pissarides (2007). Structural change in a multisector model of growth. *American economic review 97*(1), 429–443.

Nordhaus, W. and P. Sztorc (2013). *DICE 2013R: Introduction and User's Manual*.

Nordhaus, W. D. . (1994). *Managing the Global Commons: The Economics of Climate Change*. MIT Press.

Nordhaus, W. D. and A. Moffat (2017). A survey of global impacts of climate change: Replication, survey methods, and a statistical analysis. Working Paper 31323, National Bureau of Economic Research.

Obstfeld, M. (1984). Multiple stable equilibria in an optimizing perfect-foresight model. *Econometrica 52*(1), 223–228.

Obstfeld, M. and K. Rogoff (1983). Speculative hyperinflations in maximizing models: Can we rule them out?. *Journal of Political Economy 91*(4), 675 – 687.

Obstfeld, M. and K. S. Rogoff (1995). Exchange rate dynamics redux. *Journal of Political Economy 103*(3), 624–60.

Obstfeld, M. and K. S. Rogoff (1996, December). *Foundations of International Macroeconomics*, Volume 1 of *MIT Press Books*. The MIT Press.

Obstfeld, M. and K. S. Rogoff (2001, None). The six major puzzles in international macroeconomics: Is there a common cause?

Ohanian, L., A. Raffo, and R. Rogerson (2008). Long-term changes in labor supply and taxes: Evidence from oecd countries, 1956–2004. *Journal of Monetary Economics 55*(8), 1353–1362.

Ohanian, L. E., M. Orak, and S. Shen (2023). Revisiting capital-skill complementarity, inequality, and labor share. *Review of Economic Dynamics 51*, 479–505.

Olea, J. L. M., M. Plagborg-Møller, E. Qian, and C. K. Wolf (2024). Double robustness of local projections and some unpleasant varithmetic. National Bureau of Economic Research working paper 32495.

Olsson, J. et al. (2019). Structural transformation of the labor market and the aggregate economy. *Unpublished Manuscript, University of Amsterdam*.

OPEC (2019). Opec annual statistical bulletin. Online annual statistical bulletin 2019.

Ottonello, P. and D. J. Perez (2019). The currency composition of sovereign debt. *American Economic Journal: Macroeconomics 11(3)*, 174–208.

Ozkan, S., J. Hubmer, S. Salgado, and E. Halvorsen (2023). Why are the wealthiest so wealthy? a longitudinal empirical investigation.

Parente, S. L. and E. C. Prescott (1994). Barriers to technology adoption and development. *Journal of political Economy 102*(2), 298–321.

Parker, J. A., N. S. Souleles, D. S. Johnson, and R. McClelland (2013). Consumer spending and the economic stimulus payments of 2008. *American Economic Review 103*(6), 2530–2553.

Peters, M. and C. Walsh (2023). Population growth and firm dynamics. *Journal of Political Economy Macroeconomics*.

Phelps, E. S. (1967). Phillips curves, expectations of inflation and optimal unemployment over time. *Economica 34*(135), 254–281.

Phillips, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the united kingdom, 1861-1957. *economica 25*(100), 283–299.

Piazzesi, M. and M. Schneider (2016). Housing and macroeconomics. *Handbook of macroeconomics 2*, 1547–1640.

Pigou, A. C. . (1920). *The economics of welfare*. Macmillan.

Pijoan-Mas, J. (2006). Precautionary savings or working longer hours? *Review of Economic Dynamics 9*, 326–352.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Cambridge, MA: The Belknap Press of Harvard University Press.

Pissarides, C. A. (1985a). Short-run equilibrium dynamics of unemployment vacancies, and real wages. *American Economic Review 75*(4), 676–90.

Pissarides, C. A. (1985b). Short-run equilibrium dynamics of unemployment, vacancies, and real wages. *American Economic Review 75*, 676–690.

Pissarides, C. A. (2000). *Equilibrium Unemployment Theory, 2nd ed.* Cambridge, Massachusetts: MIT Press.

Pistaferri, L. (2003). Anticipated and unanticipated wage changes, wage risk, and intertemporal labor supply. *Journal of Labor Economics 21*(3), 729–754.

Plagborg-Møller, M. and C. K. Wolf (2021). Local projections and vars estimate the same impulse responses. *Econometrica 89*(2), 955–980.

Prescott, E. C. (1986). Theory ahead of business cycle measurement. *Quarterly Review* (Fall), 9–22.

Prescott, E. C. (2004). Why do americans work so much more than europeans? *Federal Reserve Bank of Minneapolis Quarterly Review 28*(1), 2–15.

Quadrini, V. (2000). Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics 3*(1), 1–40.

Quadrini, V. and J.-V. Ríos-Rull (2015). Inequality in macroeconomics. In *Handbook of Income Distribution*, Volume 2, pp. 1229–1302. Elsevier.

Rabanal, P. and J. F. Rubio-Ramírez (2015). Can international macroeconomic models explain low-frequency movements of real exchange rates? *Journal of International Economics 96*(1), 199–211.

Raffo, A. (2008). Net exports, consumption volatility and international business cycle models. *Journal of International Economics 75*(1), 14–29.

Ramey, V. A. (2011). Identifying government spending shocks: It's all in the timing. *The Quarterly Journal of Economics 126*(1), 1–50.

Ramey, V. A. and N. Francis (2009, July). A century of work and leisure. *American Economic Journal: Macroeconomics 1*(2), 189–224.

Ramsey, F. P. (1927). A contribution to the theory of taxation. *ej 37*, 47–61.

Ramsey, F. P. (1928). A mathematical theory of saving. *The economic journal 38*(152), 543–559.

Ravn, M. O. and H. Uhlig (2002). On adjusting the Hodrick-Prescott filter for the frequency of observations. *The Review of Economics and Statistics 84*(2), 371–375.

Rawls, J. (1971). *A Theory of Justice: Original Edition*. Harvard University Press.

Rebelo, S. (1991). Long-run policy analysis and long-run growth. *Journal of political Economy 99*(3), 500–521.

Reinhart, C., K. S. Rogoff, and M. A. Savastano (2003). Debt intolerance. *Brookings Papers on Economic Activity*, 1–74.

Reinhart, C. M., V. Reinhart, and K. S. Rogoff (2015). Dealing with debt. *Journal of International Economics, forthcoming 96 (S1)*, 43–51.

Reinhart, C. M. and K. S. Rogoff (2009). *This Time is Different: Eight Centuries of Financial Folly*. Princeton, New Jersey: Princeton University Press.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*, 707–720.

Restuccia, D. and R. Rogerson (2017). The causes and costs of misallocation. *Journal of Economic Perspectives 31*(3), 151–174.

Rivera-Batiz, L. A. and P. M. Romer (1991). Economic integration and endogenous growth. *The Quarterly Journal of Economics 106*(2), 531–555.

Roch, F. and F. Roldán (2023). Uncertainty premia, sovereign default risk, and state-contingent debt. *Journal of Political Economy Macroeconomics 1(2)*, 334–370.

Rocheteau, G. and E. Nosal (2017). *Money, payments, and liquidity*. Cambridge MA: MIT Press.

Rogerson, R. (1988). Indivisible labor, lotteries and equilibrium. *Journal of Monetary Economics 21*(1), 3–16.

Rogerson, R. (2024). Why labor supply matters for macroeconomics. *Journal of Economic Perspectives 38*(2), 137–158.

Rogerson, R. and J. Wallenius (2009). Micro and macro elasticities in a life cycle model with taxes. *Journal of Economic Theory 144*(6), 2277–2292.

Rogner, H. (1997). An assessment of world hydrocarbon resources. *Annual Review of Energy and the Environment 10*(22), 217–262.

Romer, C. D. and D. H. Romer (1989). Does monetary policy matter? a new test in the spirit of friedman and schwartz. *NBER macroeconomics annual 4*, 121–170.

Romer, C. D. and D. H. Romer (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review 94*(4), 1055–1084.

Romer, C. D. and D. H. Romer (2023). Presidential address: Does monetary policy matter? the narrative approach after 35 years. *American Economic Review 113*(6), 1395–1423.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of political economy 94*(5), 1002–1037.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy 98*(5, Part 2), S71–S102.

Rouwenhorst, K. G. (1995). Asset pricing implications of equilibrium business cycle models. In T. F. Cooley (Ed.), *Frontiers of Business Cycle Research*, Princeton, NJ. Princeton University Press.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers 3*(2), 135–146.

Samuelson, P. (1958a). An exact consumption-loan model of interest with or without the social contrivance of money. *jpe 66*, 467–482.

Samuelson, P. (1958b). An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy 66*.

Sandmo, A. (1970). The effect of uncertainty on saving decisions. *The Review of Economic Studies 37*(3), 353–360.

Sargent, T. J. and P. Surico (2011, February). Two illustrations of the quantity theory of money: Breakdowns and revivals. *American Economic Review 101*(1), 109–28.

Sargent, T. J. and N. Wallace (1985). Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review 9*(1), 15 – 31.

Schechtman, J. and V. L. Escudero (1977). Some results on "an income fluctuation problem". *Journal of Economic Theory 16*(2), 151–166.

Schmitt-Grohé, S. and M. Uribe (2003). Closing small open economy models. *Journal of International Economics 61*(1), 163–185.

Schmitt-Grohé, S., M. Uribe, and M. Woodford (2022). *International Macroeconomics: A Modern Approach*. Princeton University Press.

Schoellman, T. (2012). Education quality and development accounting. *The Review of Economic Studies 79*(1), 388–417.

Scholz, C. M. and G. Ziemes (1999). Exhaustible resources, monopolistic competition, and endogenous growth. *Environmental and Resource Economics 13*(5), 169–185.

Schularick, M. and A. M. Taylor (2012). Credit booms gone bust: Monetary policy, leverage cycles and financial crises, 1870-2008. *102*(2), 1029–61.

Schulhofer-Wohl, S. (2011). Heterogeneity and tests of risk sharing. *Journal of Political Economy 119*(5), 925–958.

Shiller, R. J. (2000). *Irrational Exuberance*. Princeton, New Jersey: Princeton University Press.

Shimer, R. (2005). The cyclical behavior of equilibrium unemployment and vacancies. *American Economic Review 95*, 25–49.

Shimer, R. (2009). Convergence in macroeconomics: The labor wedge. *American Economic Journal: Macroeconomics 1*(1), 280–297.

Sibley, D. S. (1975). Permanent and transitory income effects in a model of optimal consumption with wage income uncertainty. *Journal of Economic Theory 11*(1), 68–82.

Sims, C. (1980a). Macroeconomics and reality. *Econometrica 48*, 1–48.

Sims, C. A. (1972). Money, income, and causality. *American Economic Review 62*(4), 540 – 552.

Sims, C. A. (1980b). Macroeconomics and reality. *Econometrica 48*(1), 1 – 48.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European economic review 36*(5), 975–1000.

Smulders, S. a. and M. de Nooij (2003). The impact of energy conservation on technology and economic growth. *Resource and Energy Economics 25*, 59–79.

Solow, R. M. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics 70*(1), 65–94.

Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, 312–320.

Solow, R. M. (1987). New york times book review. *July 12*(1987), 36.

Steinsson, J. (2003). Optimal monetary policy in an economy with inflation persistence. *Journal of Monetary Economics 50*(7), 1425–1456.

Stiglitz, J. E. and M. Rothschild (1970). Increasing risk: I. A definition. *Journal of Economic Theory 2*(3), 225–243.

Stockman, A. C. and L. L. Tesar (1995, March). Tastes and technology in a two-country model of the business cycle: Explaining international comovements. *American Economic Review 85*(1), 168–185.

Stokey, N. L. and E. C. Lucas, R. E. with Prescott (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.

Stokey, N. L. and S. Rebelo (1995). Growth effects of flat-rate taxes. *Journal of political Economy 103*(3), 519–550.

Storesletten, K., C. I. Telmer, and A. Yaron (2001). The welfare cost of business cycles revisited: Finite lives and cyclical variation in idiosyncratic risk. *European Economic Review 45*(7), 1311–1339.

Straub, L. and I. Werning (2020, January). Positive long-run capital taxation: Chamley-judd revisited. *American Economic Review 110*(1), 86–119.

Sturzenegger, F. and J. Zettelmeyer (2005). Haircuts: estimating investor losses in sovereign debt restructurings, 1998-2005. IMF Working Paper 05-137.

Syverson, C. (2025). Markups and Markdowns. *Annual Review of Economics 17*, 52–76.

Tauchen, G. (1986). Finite state markov-chain approximations to univariate and vector autoregressions. *Economics Letters 20*, 177–181.

Taylor, A. M. (2002, February). A century of purchasing-power parity. *The Review of Economics and Statistics 84*(1), 139–150.

Taylor, A. M. and M. P. Taylor (2004, December). The purchasing power parity debate. *Journal of Economic Perspectives 18*(4), 135–158.

Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of Political Economy 88*, 1–23.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy 39*, 195–214.

Tomz, M. and M. L. J. Wright (2007). Do countries default in "bad times"? *Journal of the European Economic Association 5*, 352–360.

Townsend, R. M. (1980). *Models of money with spatially separated agents*, pp. 265–303. Federal Reserve Bank of Minneapolis Minneapolis.

Townsend, R. M. (1994). Risk and insurance in village india. *Econometrica*, 539–591.

Uhlig, H. (1996). A law of large numbers for large economies. *Economic Theory 8*, 41–50.

Uhlig, H. (2001). A Toolkit for Analysing Nonlinear Dynamic Stochastic Models Easily. In R. Marimon and A. Scott (Eds.), *Computational Methods for the Study of Dynamic Economies*, Oxford, UK. Oxford University Press.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics 52*(2), 381–419.

Uribe, M. and V. Yue (2006). Country spreads and emerging countries: Who drives whom? *Journal of International Economics 69*, 6–36.

Uzawa, H. (1961, None). On a two-sector model of economic growth. *The Review of Economic Studies 29*(1), 40–47.

Wallace, N. (1981a). A hybrid fiat-commodity monetary system. *Journal of Economic Theory 25*(3), 421 – 430.

Wallace, N. (1981b). A modigliani-miller theorem for open-market operations. *The American Economic Review 71*(3), 267–274.

Wallace, N. (1998). A dictum for monetary theory. *Federal Reserve Bank of Minneapolis Quarterly Review 22*(1), 20 – 26.

Weitzman, M. L. (1974). Prices vs. quantities. *The Review of Economic Studies 41*(4), 477–491.

Whaples, R. (1990). *The Shortening of the American Work Week: An Economic and Historical Analysis of its Context, Causes, and Consequences*. Ph. D. thesis, University of Pennsylvania.

Wieland, J. F. and M.-J. Yang (2016, March). Financial dampening. Working Paper 22141, National Bureau of Economic Research.

Williamson, J. G. (1995). The evolution of global labor markets since 1830: Background evidence and hypotheses. *Explorations in Economic History 32*(2), 141–196.

Woodford, M. (2003a). *Interest and Prices–Foundations of a Theory of Monetary Policy*. Princeton and Oxford: Princeton University Press.

Woodford, M. (2003b, October). Optimal interest-rate smoothing. *Review of Economic Studies 70*(4), 861–886.

Zeldes, S. P. (1989). Optimal consumption with stochastic income: Deviations from certainty equivalence. *The Quarterly Journal of Economics 104*(2), 275–298.

Appendices

3.A Appendix to Chapter 3

This appendix proves a discrete-time version of the Uzawa (1961) Theorem, following the continuous-time proof by Schlicht (2006). Schlicht's continuous-time proof was also discussed in Jones and Scrimgeour (2008). We will separate the Theorem into two parts.

First, consider the following aggregate production function

$$Y_t = \tilde{F}_t(K_t, L_t), \tag{3.A.1}$$

where Y_t is output, K_t is capital, and L_t is labor input. The sequence of functions $\tilde{F}_t$ is defined for all K and L in the positive orthant and each $\tilde{F}_t$ exhibits constant returns to scale. Assume that the population (labor) grows at the constant rate n:

$$L_t = L_0(1+n)^t.$$

The first theorem asserts that, if there is a balanced growth path where Y_t and K_t grow at the same rate γ, the sequence of functions $\tilde{F}_t$ has to take a particular form, under which technological progress is labor-augmenting (Harrod neutral).

Theorem .1 (Uzawa Theorem, Part I) *Suppose that for all Y, K, and L such that $Y = F_0(K, L)$,*

$$Y_t = \tilde{F}_t(K_t, L_t), \text{ where } Y_t = Y(1+\gamma)^t, K_t = K(1+\gamma)^t, \text{ and } L_t = L(1+n)^t$$

holds for all t. That is, the sequence of functions $\tilde{F}_t$ is consistent with balanced growth, regardless of the initial conditions and factor combinations. Then there exists a function $F(K, L)$, defined for all K and L, such that for all K_t, L_t, and t

$$\tilde{F}_t(K_t, L_t) = F(K_t, A_t L_t).$$

In particular, the function F is identical to $\tilde{F}_0$. The variable A_t grows at a constant rate $(1+\gamma)(1+n) - 1$.

Proof. Take any K_t, L_t, and t. Define K by $K = K_t/(1+\gamma)^t$, L by $L = L_t/(1+n)^t$, and Y by $Y = \tilde{F}_t(K_t, L_t)/(1+\gamma)^t$. Define $F(k, l) = \tilde{F}_0(k, l)$ for any (k, l).

We know that $Y = \tilde{F}_0(K, L)$. So $Y = F(K, L)$. Because $\tilde{F}_0$ exhibits constant returns to scale, we can multiply both sides by $(1+\gamma)^t$ and obtain $Y(1+\gamma)^t = F(K(1+\gamma)^t, L(1+\gamma)^t)$. Transferring back using the definitions of Y, K, and L, we obtain

$$\tilde{F}_t(K_t, L_t) = F(K_t, A_t L_t),$$

where

$$A_t = \left(\frac{1+\gamma}{1+n}\right)^t.$$

■

The proof makes it clear that $F(K_t, A_t L_t)$ completely characterizes $\tilde{F}_t$ over its domain.

Now, in Part II of the Theorem, we specify a setting that is common in macroeconomic modeling and show that, indeed, output and capital have to grow at the same rate when

the growth rates of capital, output, consumption, investment are constant over time. The capital accumulation equation is specified as

$$K_{t+1} - K_t = I_t - \delta K_t. \tag{3.A.2}$$

I_t is investment and $\delta > 0$ is depreciation rate. In the goods market, output equals consumption C_t plus investment I_t:

$$Y_t = C_t + I_t. \tag{3.A.3}$$

Note that the setting is more general than the Solow model covered in the main text: here, we don't impose any assumption on investment.

Theorem .2 (Uzawa Theorem, PartII) *Suppose that, under the assumptions* (3.A.2) *and* (3.A.3)*, there exists a growth path in which the investment is strictly positive* $I_t > 0$ *and the growth rates of* Y_t*,* C_t*,* I_t*, and* K_t *are constant over time. Call these growth rates* γ_Y*,* γ_C*,* γ_I*, and* γ_K*. Then* $\gamma_Y = \gamma_K$*.*

Proof. From (3.A.2),

$$1 + \gamma_K = \frac{I_t}{K_t} - \delta$$

holds. Thus, I_t/K_t has to remain constant, which implies $\gamma_I = \gamma_K$. Therefore, it suffices to show $\gamma_Y = \gamma_I$. Subtracting (3.A.3) for time t from (3.A.3) for time $t+1$:

$$C_{t+1} - C_t + I_{t+1} - I_t = Y_{t+1} - Y_t.$$

Then

$$\frac{C_{t+1} - C_t}{C_t} C_t + \frac{I_{t+1} - I_t}{I_t} I_t = \frac{Y_{t+1} - Y_t}{Y_t} Y_t$$

and thus

$$\gamma_C C_t + \gamma_I I_t = \gamma_Y Y_t$$

holds. Again using (3.A.3),

$$(\gamma_C - \gamma_Y) C_t = (\gamma_Y - \gamma_I) I_t. \tag{3.A.4}$$

Suppose, by contradiction, $\gamma_Y \neq \gamma_I$. Because $I_t > 0$, (3.A.4) implies $C_t \neq 0$ and (3.A.4) can be rewritten as

$$\frac{I_t}{C_t} = \frac{\gamma_C - \gamma_Y}{\gamma_Y - \gamma_I}$$

and therefore I_t/C_t has to remain constant. This fact means $\gamma_I = \gamma_C$, but because of (3.A.3) $\gamma_I = \gamma_C$ implies $\gamma_I = \gamma_C = \gamma_Y$. Contradiction. ■

Note that, in the Solow model, the latter half of the proof is unnecessary because $\gamma_Y = \gamma_I$ by the assumption of the constant saving rate.

4.A Appendix to Chapter 4

This Appendix includes the proofs of theorems and propositions in Chapter 4, and analyzes the NGM with a phase diagram.

4.A.1 Constraints in the consumption-saving problem

Consider the infinite horizon version of the consumption-saving problem P2, where the consumer maximizes

$$\max_{\{c_t,a_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t u(c_t),$$

subject to the sequential budget constraint

$$a_{t+1} = (1+r)a_t + w_t - c_t. \tag{4.A.1}$$

As we saw in the main text, this constraint is not sufficient in guaranteeing a well-behaved maximization problem, as Ponzi schemes are possible. In the finite-horizon model, we imposed the terminal condition $a_{T+1} \geq 0$. Would it make sense to impose a similar condition in the limit of this environment, $\lim_{T\to\infty} a_{T+1} \geq 0$? The answer is no, because such constraint would be "too restrictive." To see this, consider a situation where $w_t = \bar{w} > 0$ for all t. The consumer could borrow a small amount, for example $\varepsilon < \bar{w}/(1+r)$, which can be repaid with her income $\bar{w}$ in the following period. The outcome of this, $a_{t+1} = -\varepsilon$ for all t, violates the condition $\lim_{T\to\infty} a_{T+1} \geq 0$ but it would not constitute a Ponzi-scheme. In other words, such behavior is completely feasible in this environment. What is then the "natural borrowing limit" for this economy?

To find a reasonable alternative, let us start by assuming the flow budget constraint (4.A.1) always has to hold. Combining (4.A.1) up to time T, we can obtain

$$\sum_{t=0}^{T} \frac{c_t}{(1+r)^t} + \frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{T} \frac{w_t}{(1+r)^t}. \tag{4.A.2}$$

It seems reasonable to impose

$$\lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} \geq 0 \tag{4.A.3}$$

to ensure that

$$\sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} \leq (1+r)a_0 + \sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t} \tag{4.A.4}$$

holds as $T \to \infty$. Here, we are assuming $\sum_{t=0}^{T} w_t/(1+r)^t$ is finite. Constraint (4.A.4) corresponds to the lifetime budget constraint in the main text. Condition (4.A.3) is the no Ponzi game (nPg) condition. It is straightforward to check that imposing (4.A.3) would prevent the Ponzi scheme described in the main text.

The natural borrowing limit in this setting turns out to be

$$a_{t+1} \geq -\sum_{s=t+1}^{\infty} \frac{w_s}{(1+r)^{s-t}}. \tag{4.A.5}$$

The next Theorem shows the equivalence of different ways of imposing the constraints.

Theorem .3 *The following three constraint sets are equivalent.*

(i) *The flow budget constraint (4.A.1) for $t = 0, 1, ...$ and no Ponzi game condition (4.A.3)*

(ii) *The flow budget constraint (4.A.1) for $t = 0, 1, ...$ and the natural borrowing limit (4.A.5) for $t = 0, 1, ...$*

(iii) *The lifetime budget constraint (4.A.4)*

Proof. To show the equivalence, we start from showing that (i)⇒(ii), then (ii)⇒(iii), and then (iii)⇒(i).

(i)⇒(ii): Using (4.A.1) from time $t+1$ to T,

$$\frac{a_{T+1}}{(1+r)^T} = \frac{a_{t+1}}{(1+r)^t} + \sum_{s=t+1}^{T} \frac{w_s - c_s}{(1+r)^s} \tag{4.A.6}$$

holds. Rewriting (4.A.6):

$$a_{t+1} = (1+r)^t \frac{a_{T+1}}{(1+r)^T} - \sum_{s=t+1}^{T} \frac{w_s - c_s}{(1+r)^{s-t}}.$$

Taking $T \to \infty$,

$$a_{t+1} = (1+r)^t \lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} - \sum_{s=t+1}^{\infty} \frac{w_s - c_s}{(1+r)^{s-t}}.$$

Using (4.A.3), this equation implies

$$a_{t+1} \geq - \sum_{s=t+1}^{\infty} \frac{w_s - c_s}{(1+r)^{s-t}}.$$

Because $c_s \geq 0$ for all s,

$$a_{t+1} \geq - \sum_{s=t+1}^{\infty} \frac{w_s}{(1+r)^{s-t}}$$

holds, which is (4.A.5). Because t was arbitrary, we are done.

(ii)⇒(iii): To show that (4.A.5) for all t implies (4.A.4), first note (4.A.5) implies

$$\frac{a_{t+1}}{(1+r)^t} \geq - \sum_{s=t+1}^{\infty} \frac{w_s}{(1+r)^s}. \tag{4.A.7}$$

The flow budget constraint (4.A.1) implies (4.A.2) holds for any T:

$$\frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{T} \frac{w_t - c_t}{(1+r)^t}.$$

Thus, combining with (4.A.7),

$$(1+r)a_0 + \sum_{t=0}^{T} \frac{w_t - c_t}{(1+r)^t} \geq -\sum_{t=T+1}^{\infty} \frac{w_t}{(1+r)^t}.$$

holds. Rearranging and taking $T \to \infty$ (the right-hand side converges to zero[30]) results in (4.A.4).

(iii)⇒(i): Note that because (4.A.4) doesn't specify $a_1, a_2, ...$, one can create (4.A.2) by appropriately defining a_T. In turn, one can also create (4.A.1) that corresponds these $a_1, a_2, ...$, because (4.A.2) and (4.A.1) are equivalent.

Take a limit of (4.A.2) for $T \to \infty$,

$$\sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} + \lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t}.$$

Rewriting,

$$\lim_{T\to\infty} \frac{a_{T+1}}{(1+r)^T} = (1+r)a_0 + \sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t} - \sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t}.$$

From (4.A.4), this equation implies (4.A.3).

■

4.A.2 Balanced growth and CRRA utility

The following theorem shows that CRRA utility is the only form of utility function consistent with balanced growth. In Appendix 3.A, we have seen that capital, output, and consumption have to grow at a same rate along the balanced growth. This fact also implies that the rental rate of capital $r = f'(k)$ is constant, because $K/Y = k/f(k)$ is constant (implying that k is constant).

Theorem .4 *For a twice differentiable utility function $u(c)$ to be consistent with balanced growth with any arbitrary growth rate γ, $u(c)$ has to be a CRRA utility function:*

$$u(c) = \log(c)$$

or

$$u(c) = \frac{c^{1-\sigma}}{1-\sigma},$$

where $\sigma > 0$ and $\sigma \neq 1$.

[30] This result follows from $\sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t}$ being finite. The right-hand side can be rewritten as

$$-\sum_{t=0}^{\infty} \frac{w_t}{(1+r)^t} + \sum_{t=0}^{T} \frac{w_t}{(1+r)^t},$$

whose limit with $T \to \infty$ equals zero.

Proof. We require, for consistency with balanced growth, that

$$\frac{u'((1+\gamma)c)}{u'(c)} = \frac{1+r}{\beta} \equiv \#$$

holds for all c. The requirement must work for any γ—in which case r may end up depending on γ. Therefore, we write $\#(\gamma)$.

So we have, for each c and γ, that

$$u'((1+\gamma)c) = u'(c)\#(\gamma).$$

Because the expression holds for all c, we obtain

$$(1+\gamma)u''((1+\gamma)c) = u''(c)\#(\gamma)$$

so we can write, by dividing the second equation times c by the first,

$$\frac{(1+\gamma)cu''((1+\gamma)c)}{u'((1+\gamma)c)} = \frac{u''(c)c}{u'(c)}.$$

Given that we require the condition to hold for all γ, we know that the function u needs to be on the CRRA form: $u''(c)c/u'(c)$ must equal a constant. We label it $-\sigma$.

To see what functional form is implied, we write

$$-\frac{\sigma}{c} = \frac{u''(c)}{u'(c)}$$

and integrate so that we obtain

$$\log(c^{-\sigma}) = \log u'(c).$$

This immediate implies

$$u'(c) = c^{-\sigma} \quad \Rightarrow \quad u(c) = \frac{c^{1-\sigma}}{1-\sigma}$$

whenever $\sigma \neq 1$; when $\sigma = 1$, we obtain $u(c) = \log c$.[31] ■

4.A.3 Proof to Proposition 4.4

Consider any alternative feasible and interior sequence $\mathbf{x} \equiv \{x_{t+1}\}_{t=0}^{\infty}$, i.e., a sequence in the interior of $\Gamma(x_t)$ $\forall t$. We want to show that for any such sequence,

$$\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[\mathcal{F}\left(x_t^*,\ x_{t+1}^*\right) - \mathcal{F}\left(x_t,\ x_{t+1}\right)\right] \geq 0.$$

[31] Because of l'Hôpital's rule, we can summarize as

$$u(c) = \lim_{s\to\sigma} \frac{c^{1-s}-1}{1-s}$$

for all $\sigma \geq 0$.

Define

$$A_T(\mathbf{x}) \equiv \sum_{t=0}^{T} \beta^t \left[\mathcal{F}\left(x_t^*,\ x_{t+1}^*\right) - \mathcal{F}\left(x_t,\ x_{t+1}\right) \right].$$

We will show that, as T goes to infinity, $A_T(\mathbf{x})$ is bounded below by zero.

By concavity of $\mathcal{F}$,

$$A_T\left(\mathbf{x}\right) \geq \sum_{t=0}^{T} \beta^t \left[\mathcal{F}_1\left(x_t^*,\ x_{t+1}^*\right)\left(x_t^* - x_t\right) + \mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right)\left(x_{t+1}^* - x_{t+1}\right) \right].$$

Now notice that for each t, x_{t+1} shows up twice in the summation. Hence, we can rearrange the expression to read

$$\begin{aligned} A_T\left(\mathbf{x}\right) \ \geq \ & \sum_{t=0}^{T-1} \beta^t \left\{ \left(x_{t+1}^* - x_{t+1}\right) \left[\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta \mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) \right] \right\} + \\ & + \mathcal{F}_1\left(x_0^*,\ x_1^*\right)\left(x_0^* - x_0\right) + \beta^T \mathcal{F}_2\left(x_T^*,\ x_{T+1}^*\right)\left(x_{T+1}^* - x_{T+1}\right). \end{aligned}$$

Some information contained in the first-order conditions will now be useful:

$$\mathcal{F}_2\left(x_t^*,\ x_{t+1}^*\right) + \beta \mathcal{F}_1\left(x_{t+1}^*,\ x_{t+2}^*\right) = 0,$$

together with $x_0^* - x_0 = 0$ (x_0 can only take on one feasible value), allows us to derive

$$A_T\left(\mathbf{x}\right) \geq \beta^T \mathcal{F}_2\left(x_T^*,\ x_{T+1}^*\right)\left(x_{T+1}^* - x_{T+1}\right).$$

In addition, $\mathcal{F}_2\left(x_T^*,\ x_{T+1}^*\right) = -\beta \mathcal{F}_1\left(x_{T+1}^*,\ x_{T+2}^*\right)$, so we obtain

$$A_T\left(\mathbf{x}\right) \geq \beta^{T+1} \mathcal{F}_1\left(x_{T+1}^*,\ x_{T+2}^*\right)\left(x_{T+1} - x_{T+1}^*\right) \geq -\beta^{T+1} \mathcal{F}_1\left(x_{T+1}^*,\ x_{T+2}^*\right) x_{T+1}^*.$$

In the finite horizon case, x_{T+1}^* would have been the level of capital left out for the day after the (perfectly foreseen) end of the world; a requirement for an optimum in that case is clearly $x_{T+1}^* = 0$.

As T goes to infinity, the right-hand side of the last inequality goes to zero by the transversality condition. That is, we have shown that the utility implied by the candidate path must be higher than that implied by the alternative.

4.A.4 Analyzing the NGM using the phase diagram

The dynamics of the social planner's solution to the NGM can be summarized by the following two difference equations: the resource constraint

$$k_{t+1} - k_t = f(k_t) - \delta k_t - c_t \tag{4.A.8}$$

and the Euler equation

$$u'(c_t) = \beta u'(c_{t+1})(f'(k_{t+1}) + 1 - \delta). \tag{4.A.9}$$

The steady-state values of (k_t, c_t), denoted by $(\bar{k}, \bar{c})$, satisfy (from $k_{t+1} = k_t$ and $c_{t+1} = c_t$)

$$0 = f(\bar{k}) - \delta \bar{k} - \bar{c} \tag{4.A.10}$$

and

$$1 = \beta(f'(\bar{k}) + 1 - \delta). \tag{4.A.11}$$

From (4.A.8), we can see that $k_{t+1} > k_t$ if and only if $f(k_t) - \delta k_t - c_t > 0$, or

$$c_t < f(k_t) - \delta k_t.$$

Similarly, $k_{t+1} < k_t$ if and only if $c_t > f(k_t) - \delta k_t$ and $k_{t+1} = k_t$ if and only if

$$c_t = f(k_t) - \delta k_t. \tag{4.A.12}$$

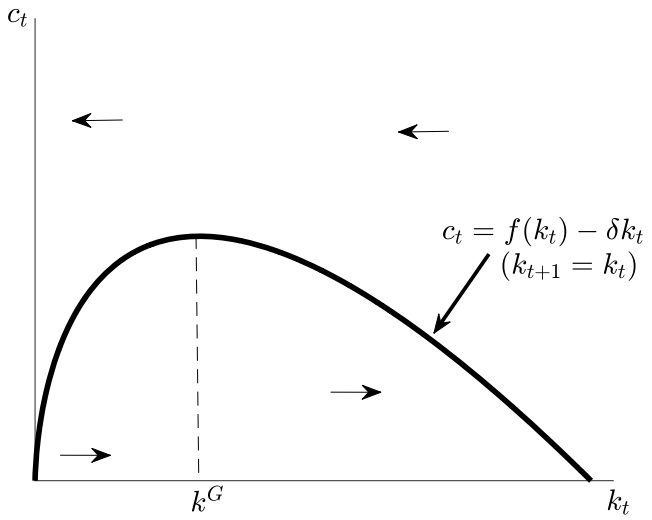


Figure .1: Drawing $k_{t+1} > k_t$, $k_{t+1} < k_t$, and $k_{t+1} = k_t$

Therefore, in the (k_t, c_t)-plane, the curve (4.A.12) divides the entire plane (in particular the positive orthant that we care) into two regions: below the curve where k_t increases over time and above the curve where k_t decreases over time. Figure .1 draws this relationship. The arrows describe how k_t moves over time in that region.

Note that it is straightforward to draw (4.A.12). The first term, $f(k_t)$, is an increasing and concave production function and the second term, δk_t, is a straight line through the origin. The vertical difference is $f(k_t) - \delta k_t$, which is hump-shaped. Note that $f(k_t) - \delta k_t$ is maximized at the point k^G where $f'(k^G) = \delta$ is satisfied. The capital stock k^G is often called the *golden rule capital stock*: it was considered as a "desirable" capital stock in earlier economic growth literature, because, when the steady-state resource constraint (4.A.10) is satisfied, this value of capital maximizes the consumption $c = f(k) - \delta k$.

From (4.A.9), $c_{t+1} > c_t$ if and only if $\beta(f'(k_{t+1}) + 1 - \delta) > 1$. Note that from (4.A.11) and because $f'(k)$ is strictly decreasing in k, this condition is equivalent to $k_{t+1} < \bar{k}$. Using (4.A.8), the condition can be rewritten as

$$c_t > f(k_t) - \delta k_t + (k_t - \bar{k}).$$

Similarly, $c_{t+1} < c_t$ if and only if $c_t < f(k_t) - \delta k_t + (k_t - \bar{k})$ and $c_{t+1} = c_t$ if and only if

$$c_t = f(k_t) - \delta k_t + (k_t - \bar{k}). \tag{4.A.13}$$

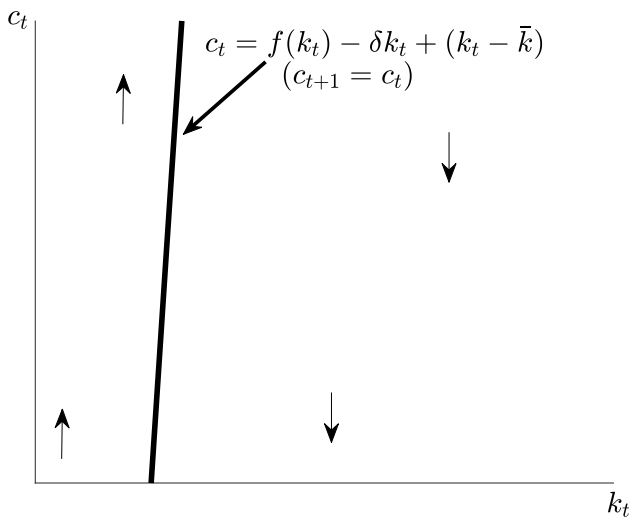


Figure .2: Drawing $c_{t+1} > c_t$, $c_{t+1} < c_t$, and $c_{t+1} = c_t$

In the (k_t, c_t) plane, the curve (4.A.13) divides the plane into two regions: above the curve, c_t increases over time, and below the curve, c_t decreases over time. Figure .2 draws this relationship. Once again, the arrows describe the movement over time. In this figure, arrows are vertical because they signify the movement of c_t.

Drawing the curve (4.A.13) is also straightforward. Because the first two terms on the right-hand side are the same as the right-hand side in (4.A.12), we simply need to add $(k_t - \bar{k})$ to the hump-shaped curve we have already drawn.

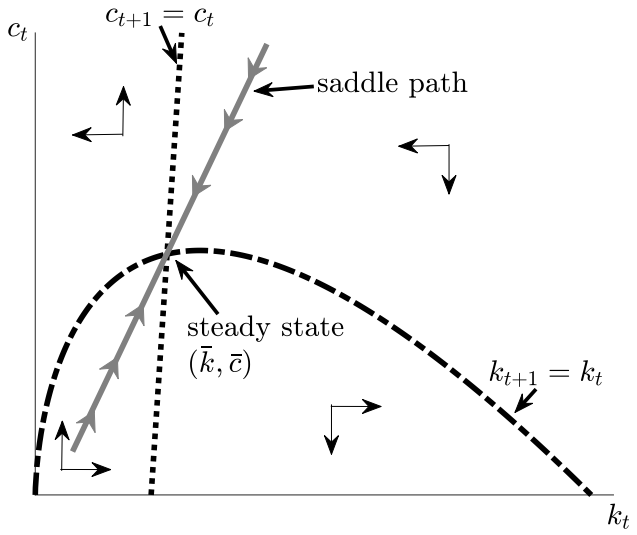


Figure .3: The phase diagram

Figure .3 puts two curves together. This diagram is called the *phase diagram*. Note here that the steady state $(\bar{k}, \bar{c})$, which corresponds to the crossing point of two curves (because

it is the point where $k_{t+1} = k_t$ and $c_{t+1} = c_t$ both hold), is placed to the left of the largest point of the hump-shaped $k_{t+1} = k_t$ curve. This comparison follows from the facts that (i) the largest point of the hump-shaped curve, k^G, satisfies $f'(k^G) = \delta$, (ii) from (4.A.11) the steady state satisfies $f'(\bar{k}) = \delta + 1 - 1/\beta$ (and the right-hand side is larger than δ), and (iii) $f'(k)$ is decreasing in k.

It turns out that the dynamics of (k_t, c_t) exhibit a *saddle-path* dynamics. The line with arrows in the figure is called the "saddle path" or the "stable arm." The (k_t, c_t) sequence converges to the steady state if and only if it starts on the saddle path. The steady state, in this case, is called the "saddle point," and the property of the dynamics is also referred to as being *saddle-path stable* (or *saddle-point stable*). If (k_t, c_t) is not on the saddle path, it will eventually diverge from the steady state.

To analyze the dynamics, recall how the economy evolves. First, for a given k_0, the social planner chooses c_0. Once c_0 is chosen, the conditions (4.A.8) and (4.A.9) determines (k_1, c_1), and the sequence of (k_t, c_t) can be determined by these difference equations. The only question is: which value of c_0 should the social planner choose? It turns out that the social planner has to choose c_0 so that (k_0, c_0) is on the saddle path. Thus, the path of (k_t, c_t) follows the saddle path and converges to the steady state.

To see why c_0 on the saddle path has to be chosen for a given k_0, first (counterfactually) suppose that c_0 is chosen above the saddle path. From the diagram, one can see that (k_t, c_t) eventually goes into the region where k_t keeps decreasing and c_t keeps increasing. In a finite time, k_t becomes zero, and at that point, it becomes impossible to follow the differential equations (4.A.8) and (4.A.9) while satisfying the constraints $k_t \geq 0$ and $c_t \geq 0$. Thus, this choice of c_0 is not consistent with the optimal conditions.

If c_0 is chosen from below the saddle path, eventually (k_t, c_t) goes into the region where both k_t and c_t keep decreasing. In particular, at a finite T, $k_t > k^G$ (and therefore $f'(k_t) < \delta$) for all $t > T$. One can show that this sequence of (k_t, c_t) violates the TVC. First note that, using (4.A.9) repeatedly,

$$\beta^t u'(c_t)(f'(k_t) + 1 - \delta)k_t = \beta^{t-1} u'(c_{t-1})k_t = u'(c_0)\frac{1}{\prod_{s=0}^{t-1}(f'(k_s) + 1 - \delta)}k_t$$

holds. The TVC requires

$$\lim_{t\to\infty} \beta^t u'(c_t)(f'(k_t) + 1 - \delta)k_t = \lim_{t\to\infty} u'(c_0)\frac{1}{\prod_{s=0}^{t-1}(f'(k_s) + 1 - \delta)}k_t$$

to be zero. However, both $u'(c_0) > 0$ and k_t is bounded from below by a strictly positive value (in particular, $k_t > k^G$ for all large t), and $\prod_{s=0}^{t-1}(f'(k_s)+1-\delta) \to 0$ because $f'(k_t)+1-\delta < 1$ for all $t > T$. Thus, $u'(c_t)(f'(k_t) + 1 - \delta)k_t$ diverges to $+\infty$, violating the TVC. And this result is intuitive: along this path, c_t is consistently low (the consumers are eating very little), despite that k_t keeps growing: a clear sign of oversaving. Eventually, the value of k_t becomes inefficiently high, to the extent that the net return from the capital at the margin $f'(k) - \delta$ is negative (this situation is often called as *dynamic inefficiency*). We discuss the issues of dynamic inefficiency further in Chapter 6 (and its Appendix).

It is also straightforward to check that the saddle path satisfies the TVC. Because the steady state $\bar{k} < k^G$, $f'(\bar{k}) > \delta$ and therefore $\prod_{s=0}^{t-1}(f'(k_s) + 1 - \delta) \to \infty$ as $t \to \infty$, implying $u'(c_t)(f'(k_t) + 1 - \delta)k_t \to 0$ as $t \to \infty$.

In sum, the optimal path of (k_t, c_t) converges to the steady state $(\bar{k}, \bar{c})$. This convergence property is qualitatively the same as in the Solow model. The difference here is that the saving rate is chosen by the social planner, and it can vary over time. The eventual steady state is derived from the social planner's optimizing behavior. One might wonder why the social planner eventually chooses a $\bar{k}$ below the golden rule k^G, which maximizes the steady-state consumption. The reason is that the consumer (and thus the social planner) discounts the future. Even though the consumption level is higher under k^G than under $\bar{k}$ after reaching the steady state, the consumer has to save extra to reach from $\bar{k}$ to k^G, and she can consume less during the transition. Because the consumer discounts the future, the cost from the short-run reduction of consumption is higher than the benefit of the eventual increase in consumption in the long run.

5.A Appendix to Chapter 5

To show that there is *aggregation* when u is a power function (with power $1-\sigma$), use the functional form to obtain

$$((1-\delta+r(K))k+w(K)-g(k,K))^{-\sigma}=$$

$$\beta\left(g(k,K)(1-\delta+r(G(K)))+w(G(K))-g(g(k,K),G(K))\right)^{-\sigma}\left[1-\delta+r(G(K))\right].$$

Raising both sides to $-(1/\sigma)$ we obtain

$$(1-\delta+r(K))k+w(K)-g(k,K)=$$

$$\left(\beta\left[1-\delta+r(G(K))\right]\right)^{-\frac{1}{\sigma}}\left(g(k,K)(1-\delta+r(G(K)))+w(G(K))-g(g(k,K),G(K))\right).$$

By collecting terms and defining $A(K)$, $B(K)$, $C(K)$, and $D(K)$ appropriately, we see that we need that, for all (k,K),

$$A(K)+B(K)k+C(K)g(k,K)+D(K)g(g(k,K),G(K))=0.$$

This makes clear that a solution that is linear in k given K, is available, i.e.,

$$g(k,K)=\mu(K)+\lambda(K)k,$$

when, for all K, $\mu(K)$ and $\lambda(K)$ solve

$$A(K)+C(K)\mu(K)+D(K)\left(\mu(G(K))+\lambda(G(K))\mu(K)\right)=0$$

$$B(K)+C(K)\lambda(K)+D(K)\lambda(G(K))\lambda(K)=0.$$

The first of these equations makes sure the object multiplying k is zero; the second makes sure the remainder is zero too. Thus, the Euler equation is satisfied for all (k,K).

These last two functional equations in λ and μ need to be solved and the solution (their functional forms) clearly depends on the functions taken as given here—r, w, and G—as well as on the primitive constants. For the consumer's problem, a solution can be sought independently of the shape of G. An equilibrium furthermore involves the consistency requirement that $G(K)=\mu(K)+\lambda(K)K$ for all K.

Aggregation obtains not only in this case. It also works for exponential utility and quadratic utility. Moreover, in all these three functional-form cases, it works also when one replaces consumption as an argument with an affine function of consumption, e.g., $\log(4c-3)$ or $-e^{-9c+7}$. You can see how applying these three functions, with their affine extensions, will deliver the same kind of equation in $g(k,K)$ as above. This class of preferences, in its entirety, is sometimes referred to as HARA preferences.[32]

Let us finally consider our typical closed-form example: utility is logarithmic, the production function is Cobb-Douglas, and there is full depreciation: $\delta=1$. Then $r(K)=\alpha AK^{\alpha-1}$ and $w(K)=(1-\alpha)AK^{\alpha}$. We know from before that the equilibrium law of motion will be

[32] HARA stands for *hyperbolic absolute risk aversion*.

$K' = G(K) = \alpha\beta AK^{\alpha}$. In this case, it is straightforward—but somewhat tedious—to verify that

$$k' = g(k, K) = \beta r(K)k,$$

i.e., $\mu(K) = 0$ and $\lambda(K) = \beta r(K)$. How does one find this solution? The best approach is typically to solve a 2-period economy, starting in period 2 and working backwards. Then functional forms appear and one can guess on decision rules of this sort. After substitution of the guess into the Euler equation one can then verify that the guess works, which involves finding the specific parameters of the adopted functional form as a final fixed-point problem.

6.A Appendix to Chapter 6

Let us denote $F(k, \omega_y + \omega_o) + 1 - \delta$ by $f(k)$.

Theorem .5 *A steady state k^* is efficient if and only if $R^* \equiv f'(k^*) \geq 1$.*

Intuitively, the steady state consumption is $c^* = f(k^*) - k^*$. Figure .1 shows the attainable levels of steady state capital stock and consumption $(k^*,\ c^*)$, given the assumptions on f. The $\left(k^G,\ c^G\right)$ locus corresponds to the "golden rule" level of steady state capital and consumption, that maximize c^G.

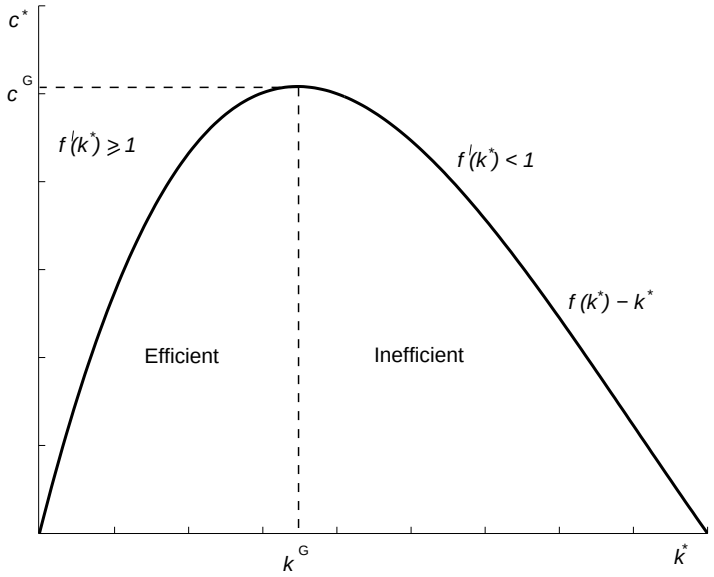


Figure .1: Efficiency of the steady state

Proof.

(i) $\underline{R^* < 1\text{: } k^* \text{ is inefficient.}}$

Assume that k^* is such that $f'(k^*) < 1$. Let c^* denote the corresponding level of steady state consumption, let $c_0 = c^*$. Now consider a change in the consumption path, whereby k_1 is set to $k_1 = k^* - \varepsilon$ instead of $k_1 = k^*$. Notice this implies an increase in c_0. Let $k_t = k_1\ \forall t \geq 1$. We have that

$$\begin{aligned} c_1 - c^* &= f(k_1) - k_1 - f(k^*) + k^* \\ &\equiv f\left(k^* - \varepsilon\right) - \left(k^* - \varepsilon\right) - f(k^*) + k^*. \end{aligned}$$

Notice that strict concavity of f implies that

$$f(k^*) < f(k^* - \varepsilon) + \left[k^* - (k^* - \varepsilon)\right] f'\left(k^* - \varepsilon\right)$$

for $\varepsilon \in \left(0,\ k^* - k^G\right)$, and we have that $f'\left(k^* - \varepsilon\right) < 1$. Therefore,

$$f(k^*) < f(k^* - \varepsilon) + k^* - \left(k^* - \varepsilon\right).$$

This implies that

$$c_1 - c^* > 0,$$

which shows that a permanent increase in consumption is feasible.

(ii) $\underline{R^* \geq 1\text{: } k^* \text{ is efficient.}}$

Suppose not, then we could decrease the capital stock at some point in time and achieve a permanent increase in consumption (or at least increase consumption at some date without decreasing consumption in the future). Let the initial situation be a steady state level of capital $k_0 = k^*$ such that $f'(k^*) \geq 1$. Let the initial c_0 be the corresponding steady state consumption: $c_0 = c^* = f(k^*) - k^*$. Since we suppose that k^* is inefficient, consider a decrease of capital accumulation at time 0: $k_1 = k^* - \varepsilon_1$, thereby increasing c_0. We need to maintain the previous consumption profile c^* for all $t \geq 1$: $c_t \geq c^*$. This requires that

$$c_1 = f(k_1) - k_2 \geq f(k^*) - k^* = c^*,$$

$$k_2 \leq f(k_1) - f(k^*) + k^*,$$

$$\underbrace{k_2 - k^*}_{\varepsilon_2} \leq f(k_1) - f(k^*).$$

Concavity of f implies that

$$f(k_1) - f(k^*) < f'(k^*)\underbrace{[k_1 - k^*]}_{-\varepsilon_1}.$$

Notice that $\varepsilon_2 \equiv k_2 - k^* < 0$. Therefore, since $f'(k^*) \geq 1$ by assumption, we have that

$$|\varepsilon_2| > |\varepsilon_1|.$$

The size of the decrease in capital accumulation is increasing. By induction, $\{\varepsilon_t\}_{t=0}^{\infty}$ is a decreasing sequence (of negative terms). Since it is bounded below by $-k^*$, we know from real analysis that it must have a limit point $\varepsilon_\infty \in [-k^*, 0)$. Consequently, the consumption sequence converges as well:

$$c_\infty = f(k^* - \varepsilon_\infty) - (k^* - \varepsilon_\infty).$$

It is straightforward to show, using concavity of f, that

$$c_\infty < c^*.$$

Then the initial increase in consumption is not feasible if the restriction is to maintain at least c^* as the consumption level for all the remaining periods of time.

■

7.A Appendix to Chapter 7

7.A.1 Recursive equilibrium for the stochastic growth model

When defining a recursive equilibrium, the first challenge is to identify the aggregate state variable(s). In the neoclassical growth model without uncertainty, we saw that the aggregate state was the stock of capital. If there are stochastic shocks to productivity, and we consider such shocks here, some aspect of the productivity process will need to be added to the aggregate state. What precise aspect depends on the nature of the productivity process. If we assume that it is first-order Markov in nature, so that nothing beyond the current value of ω will matter either for productivity today or for the probabilities of different outcomes of ω in the future, we can simply add ω; hence the aggregate state variable is (K, ω).

Definition 16 *A **recursive competitive equilibrium** consists of functions $r(K,\omega)$, $w(K,\omega)$, $G^\star(K,\omega)$, $V^\star(k,K,\omega)$, and $g^\star(k,K,\omega)$ such that*

1. *$V^\star(k,K,\omega)$ solves, for all (k,K,ω),*

   $$V(k,K,\omega) = \max_{k'} u((1-\delta+r(K,\omega))k + w(K,\omega) - k') + \beta V(k', G^\star(K,\omega), \omega') \quad \forall (k,K).$$

   and $k' = g^\star(k,K,\omega)$ attains the maximum in this problem

2. *for all K, $r(K,\omega) = A(\omega)F_K(K,1)$ and $w(K,\omega) = A(\omega)F_L(K,1)$; and*

3. *$G^\star(K,\omega) = g^\star(K,K,\omega)$ for all (K,ω).*

7.A.2 Proof of the law of iterated expectations

For any $\tau \geq t \geq 0$

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^t)\right]\right\} = \mathbb{E}_0\left[x_\tau(\omega^\tau)\right].$$

This is an application law of iterated expectations, which is more general in that it does not just apply to stochastic processes. To prove this, note

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} = \sum_{\omega^t} \pi\left(\omega^t\right) \mathbb{E}_\tau\left[x_\tau(\omega^\tau)\right].$$

Now replace the conditional expectation $\mathbb{E}_\tau$

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} = \sum_{\omega^t} \pi\left(\omega^t\right) \sum_{\omega^\tau} x_\tau(\omega^\tau)\pi(\omega^\tau|\omega^t).$$

Now notice that $\pi\left(\omega^t\right)\pi(\omega^\tau|\omega^t) = \pi(\omega^t \cup \omega^\tau) = \pi(\omega^t|\omega^\tau)\pi(\omega^\tau)$. Because $t < \tau$, once we know ω^τ we already have observed ω^t. This means that the probability distribution $\pi(\omega^t|\omega^\tau)$ is degenerate—it places all the probability mass on a single ω^t. Continuing, we have

$$\begin{aligned}\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} &= \sum_{\omega^t}\sum_{\omega^\tau} x_\tau(\omega^\tau)\pi\left(\omega^\tau\right)\pi(\omega^t|\omega^\tau) \\ &= \sum_{\omega^\tau} x_\tau(\omega^\tau)\pi\left(\omega^\tau\right)\sum_{\omega^t}\pi(\omega^t|\omega^\tau),\end{aligned}$$

where the second line follows from changing the order of summation and noting that $x_\tau(\omega^\tau)\pi\left(\omega^\tau\right)$ does not depend on t. Finally, we use the fact that the probability distribution $\pi(\omega^t|\omega^\tau)$ sums to one so we have

$$\mathbb{E}_0\left\{\mathbb{E}_t\left[x_\tau(\omega^\tau)\right]\right\} = \sum_{\omega^\tau} x_\tau(\omega^\tau)\pi\left(\omega^\tau\right) = \mathbb{E}_0\left[x_\tau(\omega^\tau)\right].$$

8.A Appendix to Chapter 8

8.A.1 Steady state of the model

The stochastic processes give us $\bar{A} = \bar{\eta} = 1$. The Euler equation gives us $\bar{r} = \beta^{-1} - 1 + \delta$. The rental rate is equal to the marginal product of capital so we have $\bar{y}/\bar{k} = \bar{r}/\alpha$. We now divide the production function by $\bar{k}$ to obtain $\bar{k}/\bar{\ell} = (\bar{y}/\bar{k})^{1/(\alpha-1)}$. Taking the product of the last two equations we have

$$\frac{\bar{y}}{\bar{\ell}} = \frac{\bar{y}}{\bar{k}} \times \frac{\bar{k}}{\bar{\ell}} = \left(\frac{\bar{r}}{\alpha}\right)^{\alpha/(\alpha-1)}.$$

As the marginal product of labor is equal to the wage we have $\bar{w} = (1-\alpha)\bar{y}/\bar{\ell}$. We can now reformulate the labor supply condition, government spending rule, and resource constraint into a system of three equations in the unknowns $\bar{c}$, $\bar{\ell}$, and $\bar{G}$

$$\begin{aligned}
\bar{c} &= \bar{w}\bar{\ell}^{-\psi} \\
\bar{G} &= (\bar{\eta}\bar{c}\gamma)^{-\gamma_1} \\
\frac{\bar{y}}{\bar{\ell}}\bar{\ell} &= \bar{c} + \delta\frac{\bar{k}}{\bar{\ell}}\bar{\ell} + \bar{G}.
\end{aligned}$$

To solve this system, we can substitute the first two equations into the third to obtain one equation in $\bar{\ell}$.

8.A.2 The fiscal multiplier in the static model

In steady state, $\bar{\eta} = 1$ so (8.3) becomes $\bar{G}/\bar{y} = \gamma/(1+\gamma)$. The total derivative of (8.2) evaluate at steady state is

$$dy_t = \frac{\bar{y}}{\bar{A}}dA_t + \frac{1}{1+\psi}\left(1 - \frac{\bar{G}}{\bar{y}}\right)^{-1}\frac{\bar{G}}{\bar{y}}\left[\frac{\bar{y}}{\bar{G}}dG_t - dy_t\right].$$

We then solve for dy_t

$$dy_t = (1+\chi)^{-1}\left(\frac{\bar{y}}{\bar{A}}dA_t + \chi\frac{\bar{y}}{\bar{G}}dG_t\right),$$

where $\chi \equiv \frac{1}{1+\psi}\left(1 - \frac{\bar{G}}{\bar{y}}\right)^{-1}\frac{\bar{G}}{\bar{y}}$. The fiscal multiplier is dy_t/dG_t, which is equal to $(1+\chi)^{-1}\chi\frac{\bar{y}}{\bar{G}}$. This is the coefficient on $\hat{G}$ in (8.5) scaled by $\bar{y}/\bar{G}$.

9.A Appendix to Chapter 9

This Appendix includes the proofs of theorems and propositions in Chapter 4, and analyzes the NGM with a phase diagram.

9.A.1 Natural log and exponential function

A natural log, often denoted as $\log(x)$ or $\ln(x)$, is the inverse operation of the exponential function (e^y or $\exp(y)$, where $e = 2.71828...$). That is,

$$e^{\log(x)} = x$$

and

$$\log(e^x) = x.$$

The natural log function has the following properties:

$$\log(xy) = \log(x) + \log(y),$$

$$\log(x^\alpha) = \alpha \log(x).$$

The derivatives of the natural log function and the exponential function are:

$$\frac{d}{dx} \log(x) = \frac{1}{x}$$

and

$$\frac{d}{dx} e^x = e^x.$$

9.A.2 Composite functions

When $y = f(x)$ and $z = g(y)$, the relationship between z and x is expressed by a composite function

$$z = g(f(x)).$$

The derivative of the composite function is

$$\frac{dz}{dx} = g'(f(x))f'(x),$$

where

$$g'(y) = \frac{d}{dy} g(y)$$

and

$$f'(x) = \frac{d}{dx} f(x).$$

For example, $X(t) = e^{\gamma t}$ is the composite function of $g(y) = e^y$ and $y = f(t) = \gamma t$. Thus, the derivative

$$\frac{d}{dt} X(t) = g'(f(t))f'(t) = \frac{d}{dy} e^y \frac{d}{dt}(\gamma t) = e^y \gamma = \gamma e^{\gamma t}.$$

For another example, $Z(t) = \log(X(t))$ is the composite function of $g(y) = \log(y)$ and $y = f(t) = X(t)$. Thus,

$$\frac{d}{dt}X(t) = g'(f(t))f'(t) = \frac{d}{dy}\log(y)\frac{d}{dt}X(t) = \frac{1}{y}\dot{X}(t) = \frac{\dot{X}(t)}{X(t)}.$$

9.A.3 No-Ponzi-game condition for the case where the interest rate varies over time

The logic of the no-Ponzi-game condition is the same as in Section 4.3.1: it prevents the consumer from rolling over the debt every period to the extent that repayment is never possible. To see it from the lifetime budget constraint perspective, consider the period-by-period budget constraint:

$$a_{t+1} = w_t + (1 + r_t)a_t - c_t,$$

where $t = 0, 1, ...$ Choose a period $T > 1$ and divide both sides by $\prod_{t=1}^{T}(1 + r_t)$ for $t = 1, ...$ and add up all constraints from $t = 0$ to T. We obtain

$$\frac{a_{T+1}}{\prod_{t=1}^{T}(1 + r_t)} = \sum_{s=0}^{T} \frac{w_s}{\prod_{t=1}^{s}(1 + r_t)} + (1 + r_0)a_0 - \sum_{s=0}^{T} \frac{c_s}{\prod_{t=1}^{s}(1 + r_t)}.$$

Then we can see the nPg

$$\frac{a_{T+1}}{\prod_{t=1}^{T}(1 + r_t)} \geq 0$$

guarantees

$$\sum_{s=0}^{T} \frac{c_s}{\prod_{t=1}^{s}(1 + r_t)} \leq \sum_{s=0}^{T} \frac{w_s}{\prod_{t=1}^{s}(1 + r_t)} + (1 + r_0)a_0,$$

which is the lifetime budget constraint. The same logic applies to the limit $T \to \infty$:

$$\lim_{T\to\infty} \frac{a_{T+1}}{\prod_{t=1}^{T}(1 + r_t)} \geq 0$$

guarantees

$$\sum_{s=0}^{\infty} \frac{c_s}{\prod_{t=1}^{s}(1 + r_t)} \leq \sum_{s=0}^{\infty} \frac{w_s}{\prod_{t=1}^{s}(1 + r_t)} + (1 + r_0)a_0.$$

To interpret the continuous-time case, consider the case where we divide one period into $1/\Delta$ number of intervals with time length Δ. Let the per-period interest rate r be the function of the calendar time $r(v)$, where $v = \Delta, 2\Delta, 3\Delta, ..., T$. The periods are still indexed by integers: $t = 1, ..., T/\Delta$. The calendar time v of period t is therefore $v = t\Delta$. Now we discount the T-period ahead asset by

$$\frac{a(T + \Delta)}{\prod_{t=1}^{T/\Delta}(1 + r(t\Delta)\Delta)},$$

where we are now denoting the asset at calendar time t by $a(t)$. To consider the limit $\Delta \to 0$ of $1/(\prod_{t=\Delta}^{T}(1+r_t\Delta))$, instead of taking the limit directly, first consider the natural log

$$f(\Delta, T) \equiv \log\left(\frac{1}{\prod_{t=1}^{T/\Delta}(1+r(t\Delta)\Delta)}\right) = -\log\left(\prod_{t=1}^{T/\Delta}(1+r(t\Delta)\Delta)\right).$$

Then

$$\frac{1}{\prod_{t=1}^{T/\Delta}(1+r(t\Delta)\Delta)} = e^{f(\Delta,T)}.$$

Now, because

$$-\log\left(\prod_{t=1}^{T/\Delta}(1+r(t\Delta)\Delta)\right) = -\sum_{t=1}^{T/\Delta}\log(1+r(t\Delta)\Delta) \approx -\sum_{t=1}^{T/\Delta} r(t\Delta)\Delta,$$

where the final approximation holds for $\Delta \to 0$, and thus taking a limit,

$$\lim_{\Delta\to 0} f(\Delta, T) = \lim_{\Delta\to 0} -\log\left(\prod_{t=1}^{T/\Delta}(1+r(t\Delta)\Delta)\right) \approx \lim_{\Delta\to 0} -\sum_{t=1}^{T/\Delta} r(t\Delta)\Delta = -\int_0^T r(t)dt.$$

Therefore, the continuous-time limit of the nPg is:

$$\lim_{\Delta\to 0} \frac{a(T+\Delta)}{\prod_{t=1}^{T/\Delta}(1+r(t\Delta)\Delta)} = e^{-\int_0^T r(t)dt} a(T) \geq 0.$$

With an infinite horizon, the condition becomes

$$\lim_{T\to\infty} e^{-\int_0^T r(t)dt} a(T) \geq 0,$$

as in the main text.

Another, more direct, approach is to solve the differential equation

$$\dot{a}(t) = w(t) + r(t)a(t) - c(t).$$

Multiplying by $e^{\int_t^T r(s)ds}$ on both sides after moving $r(t)a(t)$ to the left-hand side,

$$e^{\int_t^T r(s)ds}[\dot{a}(t) - r(t)a(t)] = e^{\int_t^T r(s)ds}[w(t) - c(t)].$$

Integrating both sides from 0 to T, noting

$$\frac{d}{dt}[e^{\int_t^T r(s)ds} a(t)] = e^{\int_t^T r(s)ds}[\dot{a}(t) - r(t)a(t)],$$

we obtain

$$a(T) - e^{\int_0^T r(s)ds} a(0) = \int_0^T e^{\int_t^T r(s)ds}[w(t) - c(t)]dt.$$

Multiplying by $e^{-\int_0^T r(s)ds}$ on both sides and rearranging, we obtain

$$e^{-\int_0^T r(s)ds}a(T) = \int_0^T e^{\int_t^T r(s)ds}w(t)dt + a(0) - \int_0^T e^{\int_t^T r(s)ds}c(t)dt,$$

which is the lifetime budget constraint in the continuous-time version. Thus, it is straightforward to see that, as $T \to \infty$, the nPg

$$\lim_{T\to\infty} e^{-\int_0^T r(s)ds}a(T) \geq 0$$

is necessary for guaranteeing

$$\int_0^\infty e^{\int_t^T r(s)ds}c(t)dt \leq \int_0^\infty e^{\int_t^T r(s)ds}w(t)dt + a(0).$$

9.A.4 $c(0)$ below the saddle path does not satisfy TVC

Let us go back to the social planner's problem:

$$\max_{c(t)} \quad \int_0^\infty e^{-\rho t}u(c(t))dt$$

subject to

$$\dot{k}(t) = f(k(t)) - \delta k(t) - c(t).$$

Hamiltonian:

$$H(t) = e^{-\rho t}u(c(t)) + \mu(t)(f(k(t)) - \delta k(t) - c(t)).$$

First-order conditions:

$$e^{-\rho t}u'(c(t)) = \mu(t) \tag{9.A.1}$$

and

$$\mu(t)[f'(k(t)) - \delta] + \dot{\mu}(t) = 0. \tag{9.A.2}$$

From (9.A.1),

$$\frac{\partial}{\partial t}\log(u'(c(t))) - \rho = \frac{\dot{\mu}(t)}{\mu(t)}.$$

From (9.A.2),

$$\frac{\dot{\mu}(t)}{\mu(t)} = -[f'(k(t)) - \delta].$$

Thus,

$$\frac{\partial}{\partial t}\log(u'(c(t))) = -[f'(k(t)) - (\delta + \rho)].$$

Integrating both sides from $t = 0$ to T,

$$\int_0^T \frac{\partial}{\partial t}\log(u'(c(t)))dt = -\int_0^T [f'(k(t)) - (\delta + \rho)]dt. \tag{9.A.3}$$

The left-hand side of (9.A.3) can be rewritten as $\log(u'(c(T))) - \log(u'(c(0)))$. Here, $c(0)$ is given as an initial condition. Thus, call $a \equiv \log(u'(c(0)))$. Therefore, (9.A.3) can be rewritten as

$$\log(u'(c(T))) = a - \int_0^T [f'(k(t)) - (\delta + \rho)]dt.$$

Taking the exponential of both sides,

$$u'(c(T)) = \exp(a) \exp\left(- \int_0^T [f'(k(t)) - (\delta + \rho)]dt \right).$$

Renaming $A \equiv \exp(a)$,

$$u'(c(T)) = A \exp\left(- \int_0^T [f'(k(t)) - (\delta + \rho)]dt \right). \tag{9.A.4}$$

The TVC is

$$\lim_{T \to \infty} e^{-\rho T} u'(c(T))k(T) = 0. \tag{9.A.5}$$

Plugging (9.A.4) into this condition,

$$\lim_{T \to \infty} e^{-\rho T} A \exp\left(- \int_0^T [f'(k(t)) - (\delta + \rho)]dt \right) k(T) = 0.$$

Because $A > 0$, A can be canceled out from here. Rearranging,

$$\lim_{T \to \infty} \exp\left(- \int_0^T [f'(k(t)) - \delta]dt \right) k(T) = 0. \tag{9.A.6}$$

Thus, we check (9.A.6) to see if the TVC is satisfied. If $(k(0), c(0))$ is on the saddle path, from any $k(0)$, $k(t)$ monotonically approaches the steady-state value $\bar{k}$ that satisfies $f'(\bar{k}) - (\delta + \rho) = 0$. Notice that because $\rho > 0$, $f'(\bar{k}) - \delta > 0$ holds. Thus, with some finite time $\hat{t} \geq 0$, $f(k(t)) - \delta > \varepsilon$ is satisfied for a (small) $\varepsilon > 0$ for all $t > \hat{t}$. Therefore,

$$\begin{aligned} \int_0^T [f'(k(t)) - \delta]dt &= \int_0^{\hat{t}} [f'(k(t)) - \delta]dt + \int_{\hat{t}}^T [f'(k(t)) - \delta]dt \\ &> \int_0^{\hat{t}} [f'(k(t)) - \delta]dt + \varepsilon(T - \hat{t}). \end{aligned}$$

Because the first term on the right-hand side is finite, it is clear that $\int_0^T [f'(k(t)) - \delta]dt \to \infty$ as $T \to \infty$. Thus, $\exp\left(- \int_0^T [f'(k(t)) - \delta]dt \right) \to 0$. Because $k(T) \to k^*$ as $T \to \infty$, (9.A.6) is satisfied.

Suppose $c(0)$ is below the value at the saddle path. Because (from the phase diagram) $k(T)$ remains strictly positive, $\int_0^T [f'(k(t)) - \delta]dt \to \infty$ as $T \to \infty$ is necessary in order to satisfy (9.A.6). However, from the phase diagram, we can see that after some finite time, $\bar{t}$, $k(t)$ enters a region where $f'(k(t)) - \delta < 0$ for all $t > \bar{t}$. Therefore, achieving $\int_0^T [f'(k(t)) - \delta]dt \to \infty$ is impossible. In this case, the TVC is not satisfied, and thus, the initial choice of $c(0)$ was not optimal.

Another way of looking at the TVC (9.A.5) is to focus on $u'(c(T))$. On the saddle path, $c(t)$ and $k(T)$ converge to the constant value, and thus, $u'(c(T))k(T)$ becomes constant. As $e^{-\rho T} \to 0$ as $T \to \infty$, (9.A.5) is satisfied. When $c(0)$ is below the saddle path, $k(T)$ stays away from zero and remains finite, but $u'(c(T))$ increases over time as the Euler equation requires $\dot{c}(t) < 0$ for a sufficiently large t. It turns out that $u'(c(T))$ increases at a rate faster than the speed $e^{-\rho T}$ approaches zero (and keeps increasing towards infinity), and thus, (9.A.5) is not satisfied. If $u'(c)$ remains finite as $c \to 0$, the value of consumption reaches zero in finite time. In this case, the Euler equation will not be satisfied after that point, which contradicts the optimality.

12.A Appendix to Chapter 12

Proof of balanced-growth preferences From the Euler equation of the planner

$$\frac{U_C\left(C_t, 1-H\right)}{U_C\left(C_{t+1}, 1-H\right)} = \beta\left[\left(\frac{H}{\kappa}\right)^{1-\alpha} + (1-\delta)\right] \tag{12.A.1}$$

where $\kappa \equiv K_t/Z_t$ is constant for all t. The fact that the right-hand side is constant over time implies that the elasticity of the marginal utility with respect to consumption has to be equal to a constant $\sigma > 0$ so that the numerator and the denominator grow at a constant rate, or

$$\frac{U_{CC}\left(C_t, 1-H\right) C_t}{U_C\left(C_t, 1-H\right)} = \sigma. \tag{12.A.2}$$

Integrating both sides of (12.*A*.2) gives us the candidate functional forms $U\left(C_t, 1-H\right) = \frac{C_t^{1-\sigma}}{1-\sigma} v^1\left(1-H\right) + v^2\left(1-H\right)$ for $\sigma \neq 1$ and $U\left(C_t, 1-H\right) = \log\left(C_t\right) v^1\left(1-H\right) + v^2\left(1-H\right)$, where $v^i\left(1-H\right)$ are arbitrary functional forms, both strictly decreasing and convex in H, with $v^1(1-H)$ assuming strictly positive values.

Focusing on the first of these functional forms, the intratemporal first-order condition of the social planner states that

$$\frac{C_t^{1-\sigma}}{1-\sigma} v_H^1\left(1-H\right) + v_H^2\left(1-H\right) = C_t^{-\sigma} v^1\left(1-H\right) Z_t \left(\frac{\kappa}{H}\right)^{\alpha}. \tag{12.A.3}$$

The right hand side of (12.*A*.3) grows over time at a constant rate $\left(1-\sigma\right) g$. In order for the left-hand side to grow at the same rate, it must be that $v_H^2\left(1-H\right) = 0$, thus $v^2\left(1-H\right)$ must be equal to a constant which can be normalized to zero. The derivation for the log case is similar. In sum, preferences are consistent with balanced growth if and only if belong to the class summarized in (12.37).

13.A Appendix to Chapter 13

13.A.1 Solving for the equilibrium dynamics in the model of Section 13.3

The first-order conditions of the firm problem in (13.7) yield the following expressions for the rental rate and wage rate (where we made use of $L(t) = e^{nt}h$ to arrive at the expressions):

$$\frac{\partial \pi(t)}{\partial K(t)} = 0 \Leftrightarrow R(t) = \alpha \left(\frac{A_y e^{(\gamma_y + n)t} h}{K(t)} \right)^{1-\alpha}, \quad \forall t, \tag{13.A.1}$$

$$\frac{\partial \pi(t)}{\partial L(t)} = 0 \Leftrightarrow w(t) = (1-\alpha) A_y e^{\gamma_y t} \left(\frac{K(t)}{A_y e^{(\gamma_y + n)t} h} \right)^{\alpha}, \quad \forall t. \tag{13.A.2}$$

Substituting (13.A.1) in the return condition, $r(t) = R(t) A_x e^{\gamma_x t} - \delta - \gamma_x$, gives for the interest rate

$$r(t) = A_x e^{\gamma_x t} \alpha \left(\frac{A_y e^{(\gamma_y + n)t} h}{K(t)} \right)^{1-\alpha} - \delta - \gamma_x. \tag{13.A.3}$$

The Hamiltonian corresponding to the household problem defined in (13.4) is:

$$\mathcal{H} = e^{-(\rho - n)t} \frac{c(t)^{1-\sigma} - 1}{1-\sigma} + \lambda(t) \big[\big(r(t) - n \big) a(t) + w(t) h - c(t) \big],$$

. The first-order conditions are then given by

$$\frac{\partial \mathcal{H}}{\partial c(t)} = 0 \Leftrightarrow \lambda(t) = e^{-(\rho - n)t} c(t)^{-\sigma}, \tag{13.A.4}$$

$$\frac{\partial \mathcal{H}}{\partial a(t)} = -\dot{\lambda}(t) \Leftrightarrow \frac{\dot{\lambda}(t)}{\lambda(t)} = -\big(r(t) - n \big), \tag{13.A.5}$$

$$\dot{a}(t) = \big(r(t) - n \big) a(t) + w(t) h - c(t). \tag{13.A.6}$$

Time-differentiating (13.A.4) gives an expression for $\frac{\dot{\lambda}(t)}{\lambda(t)}$. Setting it equal to the one in (13.A.5), replacing the interest rate by (13.A.3), and rearranging gives the Euler equation

$$\frac{\dot{c}(t)}{c(t)} = \frac{r(t) - \rho}{\sigma} = \frac{A_x e^{\gamma_x t} \alpha \left(\frac{A_y e^{(\gamma_y + n)t} h}{K(t)} \right)^{1-\alpha} - \delta - \gamma_x - \rho}{\sigma}. \tag{13.A.7}$$

Furthermore, (13.A.3), (13.A.2), and the asset market clearing condition, $K(t) e^{-\gamma_x t} / A_x = a(t) e^{nt}$, allow to rewrite the household budget condition as

$$\frac{\dot{K}(t)}{K(t)} = A_x e^{\gamma_x t} \left(\left(\frac{A_y e^{(\gamma_y + n)t} h}{K(t)} \right)^{1-\alpha} - e^{nt} \frac{c(t)}{K(t)} \right) - \delta. \tag{13.A.8}$$

This equation could have been derived more directly by combining the following two equations: A standard equation of the law of motion of the capital stock

$$\dot{K}(t) = X(t) - \delta K(t), \tag{13.A.9}$$

where X denotes investment, plus (intermediate) goods market clearing

$$Y(t) = e^{nt}c(t) + X(t)e^{-\gamma_x t}/A_x, \tag{13.A.10}$$

where consumption and investment are one to $A_x e^{\gamma_x t}$ substitutes. These two equations, (13.A.9) and (13.A.10), were however not mentioned in the equilibrium definition in the main text as they are implied by the household budget and clearing in all other markets. This is a manifestation of Walras's law.

Finally, the no-Ponzi game condition and the transversality condition can with the asset market clearing condition and the Euler equation be combined to a terminal condition

$$\lim_{T\to\infty}\left\{\frac{K(T)}{A_x}e^{-(\gamma_x+\rho)T}c(T)^{-\sigma}\right\} = 0. \tag{13.A.11}$$

13.A.2 Planner's problem of the model of Section 13.3

The planner's problem can be stated as follows:

$$\max_{\{c(t),K(t)\}_{t=0}^{\infty}} \int_{t=0}^{\infty} e^{-(\rho-n)t}\frac{c(t)^{1-\sigma}-1}{1-\sigma}, \ s.t. \ (13.A.8), \forall t, \ \text{with } K(0) \text{ given}, \tag{13.A.12}$$

plus some non-negativity constraints $c(t) \geq 0$ and $K(t) \geq 0$, $\forall t$. The Hamiltonian then reads

$$\mathcal{H} = e^{-(\rho-n)t}\frac{c(t)^{1-\sigma}-1}{1-\sigma} + \mu(t)\left[A_x e^{\gamma_x t}\left(K(t)^{\alpha}\left(A_y e^{(\gamma_y+n)t}h\right)^{1-\alpha} - e^{nt}c(t)\right) - \delta K(t)\right].$$

Taking first-order conditions, time differentiating the FOC on $c(t)$, equalizing $\frac{\dot{\mu}(t)}{\mu(t)}$ and rearranging yields:

$$\frac{\dot{c}(t)}{c(t)} = \frac{A_x e^{\gamma_x t}\alpha\left(\frac{A_y e^{(\gamma_y+n)t}h}{K(t)}\right)^{1-\alpha} - \delta - \gamma_x - \rho}{\sigma},$$

and the other first-order condition is (13.A.8). This system of equations is the same system that is derived as the solution to the decentralized equilibrium. The planner solution coincides with the decentralized equilibrium because there are no distortions and markets are complete.

13.A.3 Linearizing transitional dynamics around BGP in the model of Section 13.3

We can characterize the differences in detrended consumption and capital from the steady-state values as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \ \mathbf{x} = \begin{pmatrix}\tilde{c}-\tilde{c}^{\star}\\ \tilde{k}-\tilde{k}^{\star}\end{pmatrix}, \tag{13.A.13}$$

where the matrix $\mathbf{A}$ is the Jacobian evaluated at the steady-state values. Formally, we have

$$\mathbf{A} = \begin{pmatrix}\partial\dot{\tilde{c}}/\partial\tilde{c} & \partial\dot{\tilde{c}}/\partial\tilde{k}\\ \partial\dot{\tilde{k}}/\partial\tilde{c} & \partial\dot{\tilde{k}}/\partial\tilde{k}\end{pmatrix} = \begin{pmatrix}0 & -(1-\alpha)\alpha A_x(A_y h)^{1-\alpha}\left(\tilde{k}^{\star}\right)^{\alpha-2}\tilde{c}^{\star}/\sigma\\ -A_x & \rho-n-\gamma_x\alpha(1-\sigma)/(1-\alpha)-\gamma_y(1-\sigma)\end{pmatrix}, \tag{13.A.14}$$

where the partial derivatives are evaluated at $\tilde{c}^\star$ and $\tilde{k}^\star$. In order to examine the local dynamics we are interested in the two eigenvalues of the Jacobian $\mathbf{A}$. As the model has one predetermined state variable, $\tilde{k}$, and one control variable, $\tilde{c}$, we expect to see local non-oscillating saddle path stability. Therefore, according to the Blanchard-Kahn condition both eigenvalues should be real and one negative and one positive. The polynomial that gives the two eigenvectors can be written as the implicitly defined zs that solve

$$-z\left(\rho-n-\gamma_x\frac{\alpha(1-\sigma)}{1-\alpha}-\gamma_y(1-\sigma)-z\right)=(1-\alpha)\alpha\frac{A_x^2(A_yh)^{1-\alpha}\left(\tilde{k}^\star\right)^{\alpha-2}\tilde{c}^\star}{\sigma}. \quad (13.A.15)$$

The right-hand side of this equation is a positive constant. The left-hand side is an upward sloping quadratic polynomial with zeros at 0 and $\rho-n-\gamma_x\frac{\alpha(1-\sigma)}{1-\alpha}-\gamma_y(1-\sigma)$. Condition (13.8) guarantees that the second zero of the left-hand side is at a strictly positive value of z. Graphically, it then follows immediately that both roots are real and one is strictly positive, whereas the other is strictly negative. Solving the quadratic polynomial gives for the stable root $z=\frac{1}{2}\left(\rho-n-\gamma_x\frac{\alpha(1-\sigma)}{1-\alpha}-\gamma_y(1-\sigma)\right)\left[1-\sqrt{1+4\frac{(1-\alpha)\alpha\left(A_x^2(A_yh)^{1-\alpha}\left(\tilde{k}^\star\right)^{\alpha-2}\tilde{c}^\star\right)}{\sigma\left(\rho-n-\gamma_x\frac{\alpha(1-\sigma)}{1-\alpha}-\gamma_y(1-\sigma)\right)^2}}\right]$. As we can associate this with the dynamics in capital it gives us the local speed of convergence.

The speed of convergence is given by $-z$ and the half-life is $\ln(1/2)/z$ based on the solution τ to $\exp(z\cdot\tau)=1/2$. A simple calibration with $n=0.01$, $\gamma_x=0.013$, $\gamma_y=0.014$ $\alpha=1/3$, $\sigma=1$, and $\rho=0.03$ yields a half-life of about 5.5 years. We can also use the linearized system to determine the local slope of the saddle path or to simulate the local transitional dynamics.

13.A.4 Generalizations of the AK theory

The AK model can be generalized to allow labor to play a role. An example is, as mentioned in the main text, a model with physical and human capital where both can be accumulated at steady rates. Also, if Inada conditions do not hold, an economy with a reasonable looking production function with labor playing a significant role may asymptotically converge to an AK model, as in Jones and Manuelli (1990). For instance, under CES production $Y(t)=\left((AK(t))^{\frac{\epsilon-1}{\epsilon}}+L^{\frac{\epsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}}$, with $\epsilon>1$ and a large enough saving rate, the economy will asymptote an AK model with an output elasticity of capital of 1. Multi-sector versions of such models can also allow for consumption sector(s) that feature factors that cannot be accumulated and can therefore give a role to labor. For example, Rebelo (1991) generates endogenous growth in a model with a consumption sector with the production function $Y_C(t)=K_C(t)^\alpha L_C(t)^{1-\alpha}$ and an investment sector which is AK, i.e., $Y_I(t)=AK_I(t)$. This model is reminiscent of the model with investment-specific technical change in the previous section, as the relative price of investment will decline along a decentralized balanced growth path. Here the decline in the relative price comes from factor intensity differences (and growing wages relative to the rental rate) as opposed to exogenous differences in the rate of technical change. So what matters in order to generate an endogenous balanced growth path in this class of models is that the *investment sectors* are constant return to scale in factors that can be accumulated.

13.A.5 Equilibrium definition in the expanding variety model

In this economy a decentralized equilibrium is defined as a path of prices and quantities that jointly

1. solves the final producer problem (13.22) and solves for each monopolistically competitive machine producer problem (13.25).
2. solves the household problem (13.30).
3. clears labor and assets markets, i.e., $L = L_y(t) + L_r(t)$ and (13.29), where the firm value is given by (13.27).
4. solves the problem of the R&D firm (13.28) and is in line with the law of motion of machine varieties (13.26).
5. clears the final good market, i.e., $Y(t) = c(t) + \int_0^{N(t)} \psi x(\nu, t)\, d\nu$.

One of the equations above, such as final goods market clearing, can be viewed as redundant as it follows from all the other equilibirum conditions due to Walras's law.

13.A.6 Planner solution in the expanding variety model

As $\{x(\nu,t)\}_{\nu=0}^{N}(t)$ only show up in the (static) constraint (13.48) of the planner problem, we can solve this problem in two steps: First, solve for the optimal $\{x(\nu,t)\}_{\nu=0}^{N}(t)$ in a given point in time and then solve the dynamic problem in a second step. The first-order condition with respect to each $x(\nu,t)$ implies

$$AL_y(t)^{\phi} x(\nu,t)^{-\phi} - \psi = 0, \ \forall \nu, t, \tag{13.A.16}$$

and we therefore have $x(\nu,t) = \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} L_y(t)$. By substituting in this optimal solution for the $x(\nu,t)$s, we can collapse the constraints and express them as

$$L_y(t) = \frac{c(t)}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} N(t)^{-1}. \tag{13.A.17}$$

The Hamiltonian of the dynamic problem then reads

$$\mathcal{H} = e^{-\rho t} \frac{c(t)^{1-\sigma} - 1}{1-\sigma} + \lambda(t)\eta \left(N(t)L - \frac{c(t)}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} \right).$$

The first-order conditions are:

$$\frac{\partial \mathcal{H}}{\partial c(t)} = 0 \Leftrightarrow \lambda(t) = e^{-\rho t} c(t)^{-\sigma} \frac{\psi}{\eta} \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}, \tag{13.A.18}$$

$$\frac{\partial \mathcal{H}}{\partial N(t)} = -\dot{\lambda}(t) \Leftrightarrow \frac{\dot{\lambda}(t)}{\lambda(t)} = -\eta L, \tag{13.A.19}$$

$$\frac{\partial \mathcal{H}}{\partial \lambda(t)} = \dot{N}(t) \Leftrightarrow \dot{N}(t) = \eta N(t) L - \frac{\eta c(t)}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}}. \tag{13.A.20}$$

Time differentiating (13.A.18), substituting (13.A.19), and rearranging gives the Euler equation:

$$\frac{\dot{c}(t)}{c(t)} = \frac{\eta L - \rho}{\sigma}. \tag{13.A.21}$$

Finally, we can use (13.A.21) and (13.A.17) together with (13.A.20) to find the level L_y that allows for balanced growth:

$$\frac{\dot{L}_y(t)}{L_y} = \frac{\dot{c}(t)}{c(t)} - \frac{\dot{N}(t)}{N(t)} = 0 \Rightarrow \frac{\eta L - \rho}{\sigma} = \eta(L - L_y).$$

Summarizing, the growth rate and labor in final goods production as chosen by the social planner are $g^{SP} = \frac{\eta L - \rho}{\sigma}$ and $L_y^{SP} = \frac{(\sigma-1)L + \rho/\eta}{\sigma}$, respectively.

13.A.7 Planner solution with less than proportional knowledge spillovers

The planner problem is stated in footnote 36. The problem can again be split up in a static and dynamic part. The static problem gives the first-order condition

$$x(\nu, t) = \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} L_y(t).$$

Substituting this in the resource constraint (13.58), and subsequently in the law of motion of $N(t)$, we get:

$$L_y(t) = \frac{c(t)}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} N(t)^{-1} \Rightarrow \dot{N}(t) = \eta N(t)^{\varepsilon} \left(L e^{nt} - \frac{c(t)}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} N(t)^{-1} \right).$$

The Hamiltonian of the dynamic part of the planner's problem is

$$\mathcal{H} = e^{-(\rho-n)t} \frac{c(t)^{1-\sigma} - 1}{1-\sigma} + \lambda(t)\eta \left(N(t)^{\epsilon} L e^{nt} - \frac{c(t) N(t)^{\epsilon-1}}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} \right). \tag{13.A.22}$$

The first-order conditions are

$$\frac{\partial \mathcal{H}}{\partial c(t)} = 0 \Leftrightarrow \lambda(t) = e^{-(\rho-n)t} c(t)^{-\sigma} N(t)^{1-\varepsilon} \frac{\psi}{\eta} \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}, \tag{13.A.23}$$

$$\frac{\partial \mathcal{H}}{\partial N(t)} = -\dot{\lambda}(t) \Leftrightarrow \frac{\dot{\lambda}(t)}{\lambda(t)} = -\eta \left(\varepsilon N(t)^{\varepsilon-1} L e^{nt} + (1-\varepsilon) \frac{c(t) N(t)^{\varepsilon-2}}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} \right), \tag{13.A.24}$$

$$\frac{\partial \mathcal{H}}{\partial \lambda(t)} = \dot{N}(t) \Leftrightarrow \eta \left(N(t)^{\epsilon} L e^{nt} - \frac{c(t) N(t)^{\epsilon-1}}{\psi \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} \frac{\phi}{1-\phi}} \right) = \dot{N}(t). \tag{13.A.25}$$

Time differentiating (13.A.23) gives

$$\frac{\dot{\lambda}(t)}{\lambda(t)} = -\sigma\frac{\dot{c}(t)}{c(t)} + (1-\varepsilon)\frac{\dot{N}(t)}{N(t)} - (\rho - n).$$

Since $\frac{\dot{c}(t)}{c(t)}, \frac{\dot{N}(t)}{N(t)}$ have to be constant on a balanced growth path, this condition also implies that $\frac{\dot{\lambda}(t)}{\lambda(t)}$ has to be constant. Condition (13.A.25) then implies that this can only be the case if $N(t)^{\varepsilon-1}Le^{nt}$ is constant. That in turn requires $N(t)^{1-\varepsilon}$ to grow at the same rate as e^{nt}, which holds if and only if

$$\frac{\dot{N}(t)}{N(t)} = \frac{n}{1-\varepsilon}.$$

13.A.8 Additional figures

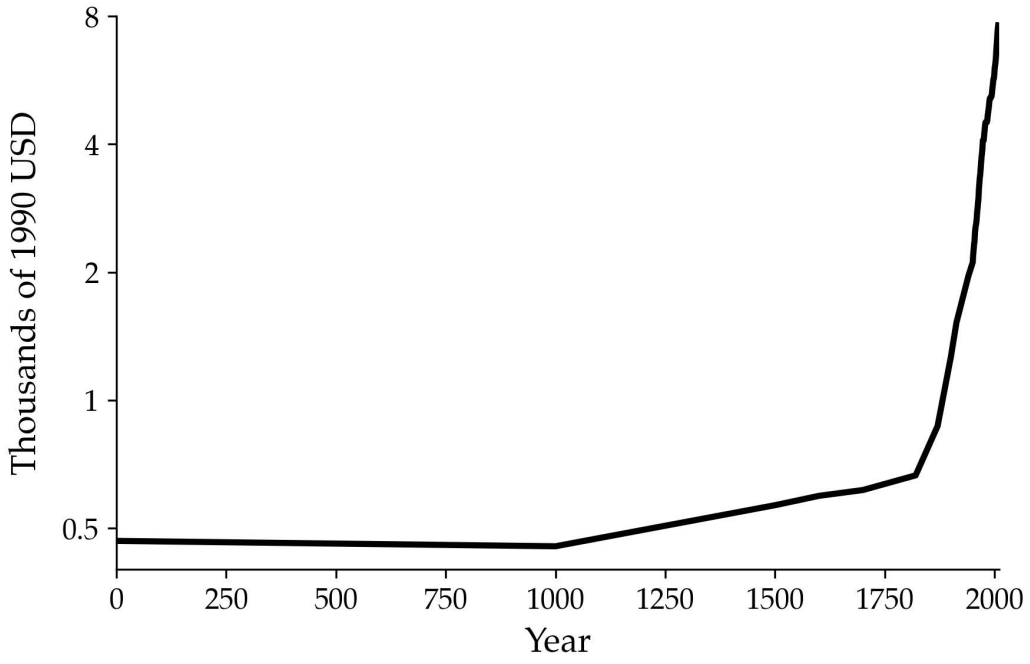


Figure 13.A.1: Global Real GDP per Capita

Notes: The data comes from the Maddison Project Database (2010).

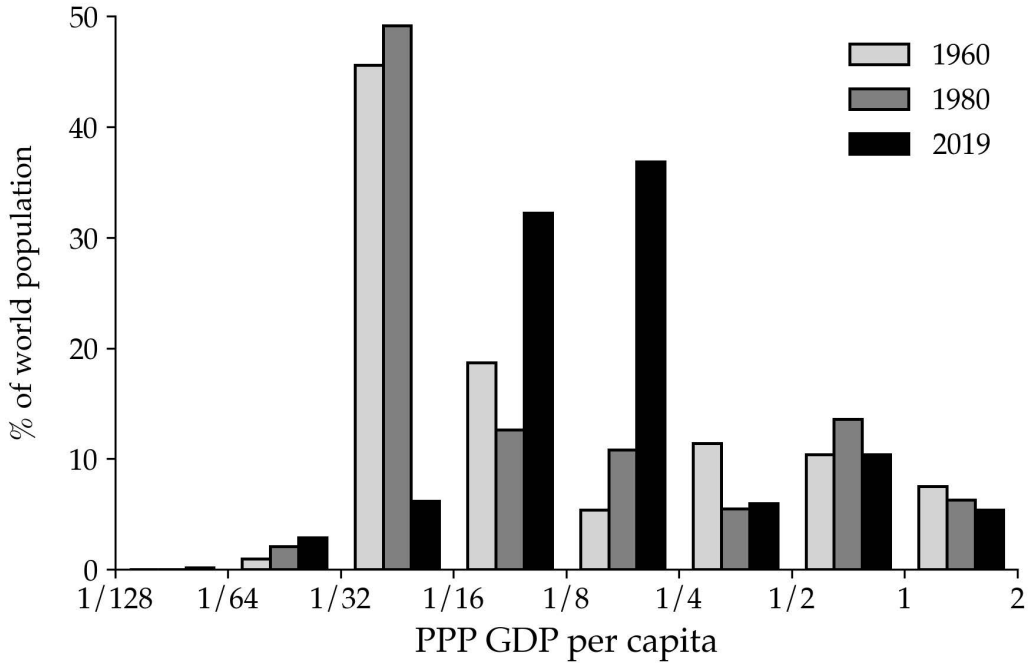


Figure 13.A.2: PPP GDP per capita in 1960, 1980 and 2019 (U.S. = 1)

Source: Penn World Table 10.0. PPP GDP per capita is calculated as the ratio of the "rgdpo" and "pop" variables relative to the U.S. in all years.

Table 13.A.1: Nominal Development Accounting in 2019

		Contributions from		
Statistic	$P_Y Y/P_C L$	$P_K K/P_Y Y$	H/L	A
Variance of log	0.98	0.11	0.08	0.69
Elasticity wrt $P_Y Y/P_C L$		0.07	0.22	0.72
90/10 ratio	11.77	1.18	1.73	5.92

Note: The data comes from the Penn World Table 10.0. Output per worker is constructed using the "rgdpo", "emp", "pl_gdpo" and "pl_c" variables. The capital to output ratio is constructed using the "cn", "rgdpo", "labsh", "pl_gdpo" and "pl_n" variables. The human capital index corresponds to the "hc" variable. We use the population-weighted average labor share of 0.53 across countries in 2019, which implies an $\alpha = 0.47$.

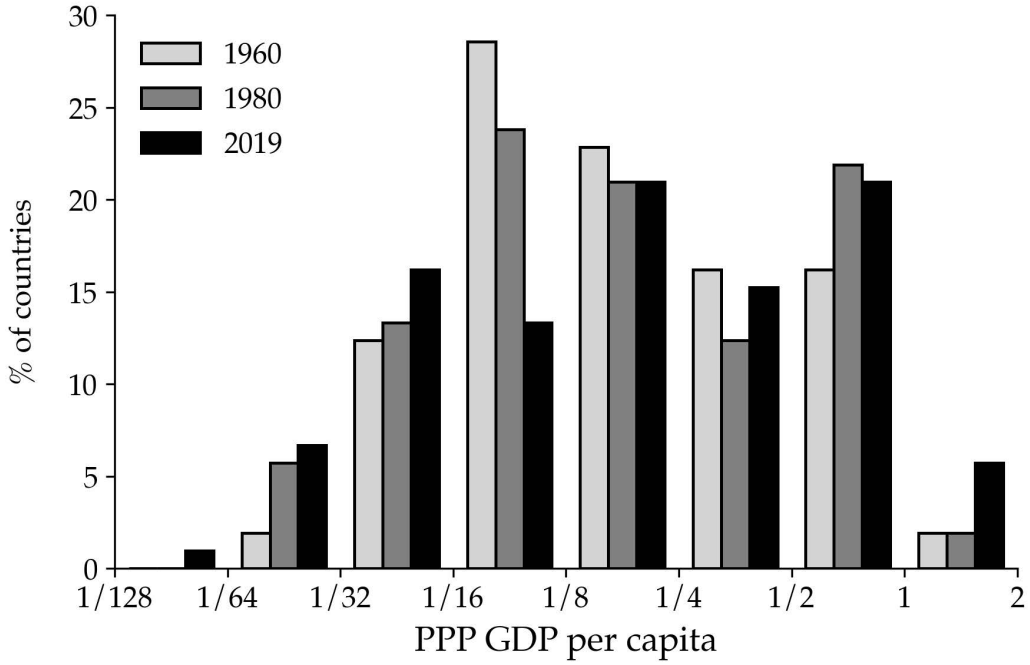


Figure 13.A.3: PPP GDP per capita in 1960, 1980 and 2019 (U.S. = 1)

Notes: The data comes from the Penn World Table 10.0. PPP GDP per capita is calculated as the ratio of the "rgdpo" and "pop" variables relative to the U.S. in all years.

Table 13.A.2: Nominal Growth Accounting for the U.S.

		Contributions from		
Period	$P_Y Y/P_C L$	$P_K K/P_Y Y$	H/L	A
1948–2020	2.07	0.09	0.38	1.60
1948–1973	3.12	-0.05	0.27	2.90
1973–1995	1.06	0.29	0.36	0.40
1995–2007	2.43	0.03	0.40	2.00
2007–2020	1.45	0.09	0.59	0.77

Note: The data comes from the U.S. Bureau of Labor Statistics. Y/L denotes real output per hour, K/Y the real capital-output ratio, H/L human capital per worker (which grows predominantly from rising years of schooling), and A is residual labor-augmenting TFP inclusive of contributions from R&D and intellectual property. The contribution of physical capital is scaled by $\alpha/(1-\alpha)$, where α is the average cost share for physical capital over the sample, equal to 0.34. See equation (13.17).

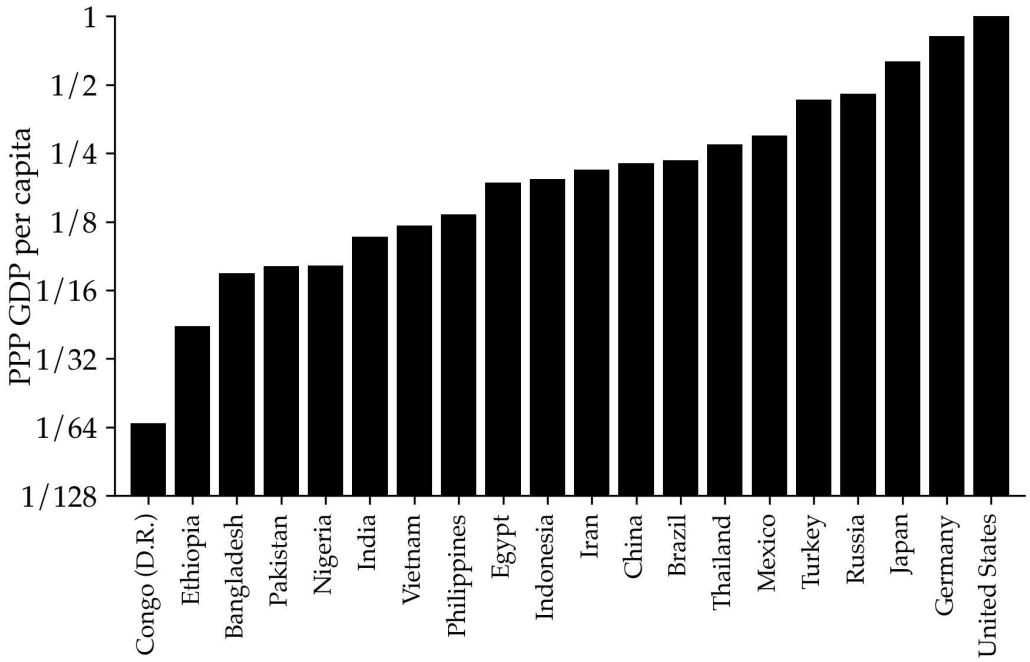


Figure 13.A.4: PPP GDP per capita in 2019 (U.S. = 1)

Source: Penn World Table 10.0. PPP GDP per capita is the ratio of the "rgdpo" and "pop" variables relative to the U.S. in 2019. We choose the largest 20 countries by population that the PWT does not classify as an outlier in terms of data quality.

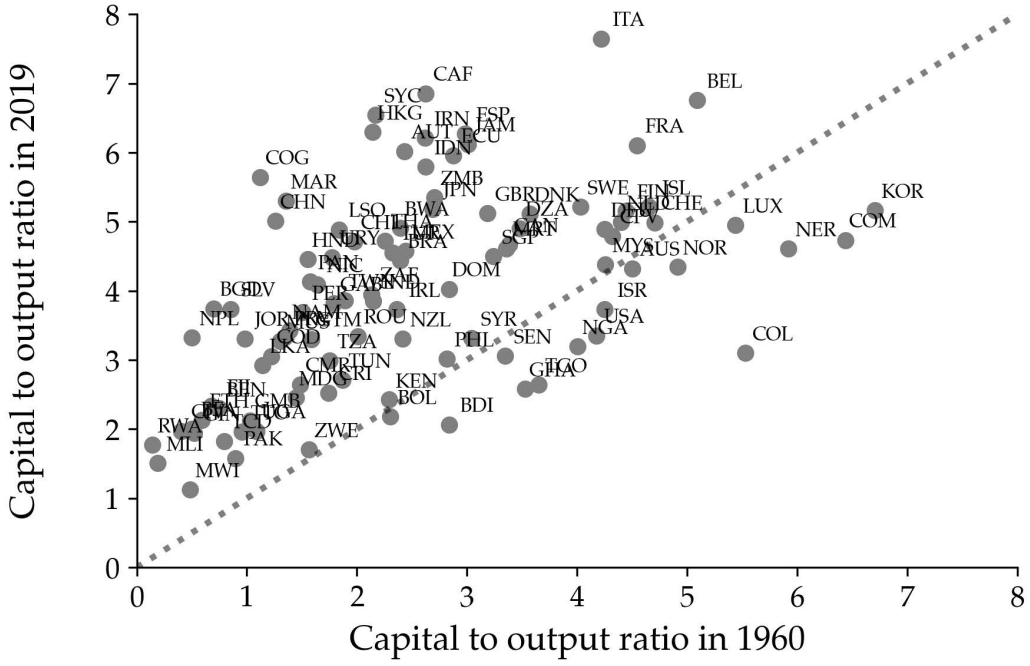


Figure 13.A.5: Capital-output ratio in 1960 vs. 2019

Notes: The data comes from the Penn World Table 10.0.

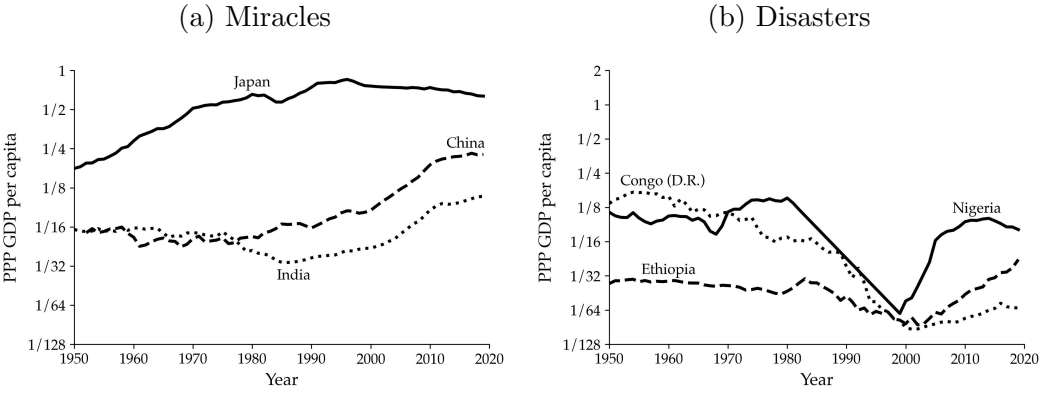


Figure 13.A.6: PPP GDP per capita for Growth Miracles and Disasters

Notes: The data comes from the Penn World Table 10.0. PPP GDP per capita is calculated as the ratio of the "rgdpo" and "pop" variables relative to the U.S. in all years.

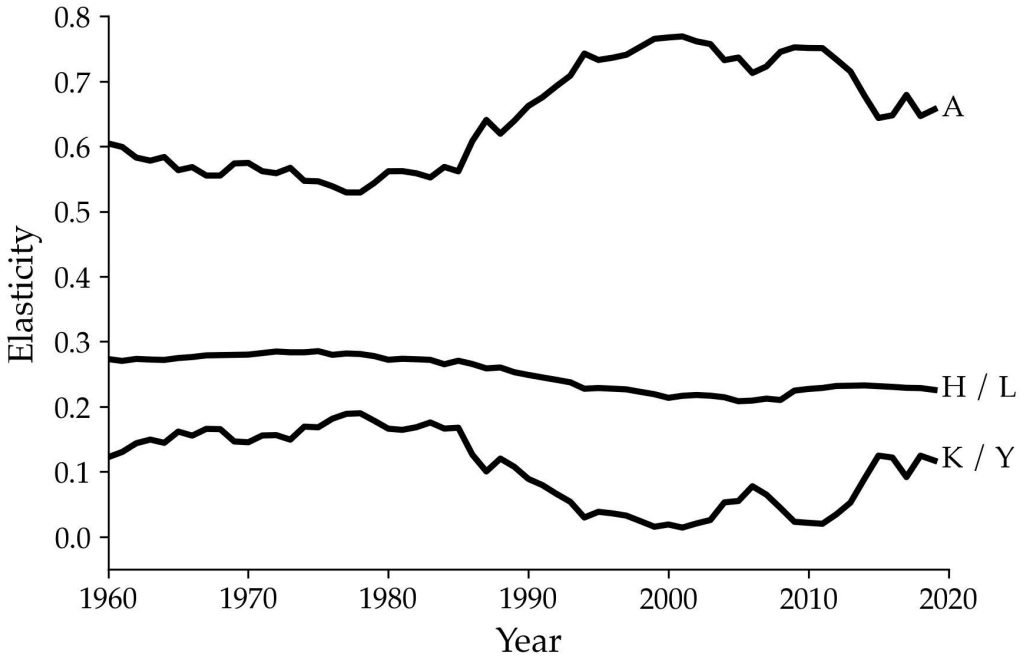


Figure 13.A.7: Elasticities

Notes: The data comes from the Penn World Table 10.0. Output per worker is constructed using the "rgdpo" and "emp" variables. The capital to output ratio is constructed using the "cn", "rgdpo" and "labsh" variables. The human capital index corresponds to the "hc" variable.

13.A.9 Drastic vs. incremental innovations

Here we illustrate how competition between different quality producers within a product line ν plays out. The final good production function can be written as $Y(t) = \frac{A}{1-\phi} L(t)^{\phi} \int_0^1 \tilde{X}(\nu,t)^{1-\phi} d\nu$ where $\tilde{X}(\nu,t) = \sum_v q(\nu,v,t)^{\frac{1}{1-\phi}} x(\nu,v,t)$ is the quality-weighted sum over all quality vintages v in line ν. If only the highest quality vintage is used, this reduces to the production function in (13.61). As the different product vintages enter production as perfect (not one-for-one) substitutes the final goods producer is indifferent between mixing any two product versions v and v' if $p(v) = \left(\frac{q(v)}{q(v')}\right)^{\frac{1}{1-\phi}} p(v')$. If instead $p(v) < \left(\frac{q(v)}{q(v')}\right)^{\frac{1}{1-\phi}} p(v')$, the producer will prefer version v over v'.

When is the highest quality product preferred over the one of second-best quality? As the highest quality $q(v)$ is a λ quality step above the second-highest quality $q(v')$ we have $q(v) = \lambda q(v')$. At constant marginal costs, and with Bertrand competition, the standard result of limit-pricing holds, i.e., the price of the second-best producer will be driven down to marginal cost. The marginal cost of producing at quality $q(v)$ is $\psi q(v)$, so the second-best producer's price is diven down to $p(v') = \psi q(v')$. Plugging that and $q(v)/q(v') = \lambda$ into the price condition, and using the definition of markup $\mu \equiv p/(\psi q)$ yields:

$$\mu \leq \lambda^{\frac{\phi}{1-\phi}}. \tag{13.A.26}$$

This gives us an upper bound for the markup a leader can charge given the competition from the second-best quality producer. From the solution of the model with 'drastic' innovations, we know that the markup of the highest-quality producer under monopolistic competition is $\mu^{\star} = \frac{1}{1-\phi}$. For this to be feasible, we must have $\mu^{\star} \leq \lambda^{\frac{\phi}{1-\phi}}$. This gives us the *drastic innovation* condition (13.64) in the main text. If, on the other hand, innovation is *incremental* and $\lambda < \left(\frac{1}{1-\phi}\right)^{\frac{1-\phi}{\phi}}$, the markup condition imposed by the second-best quality producer will bind and the markup of the leader is

$$\mu = \lambda^{\frac{\phi}{1-\phi}}. \tag{13.A.27}$$

13.A.10 The planner's solution with quality ladders

The planner problem can be split into two parts a static and a dynamic one. In the static problem the planner maximizes final output net of intermediate input production, i.e., the planner solves

$$\max_{\{x(\nu,t)\}_{\nu=0}^{1}} Y(t) - X(t) = \frac{A}{1-\phi} L^{\phi} \int_0^1 q(\nu,t) x(\nu,t)^{1-\phi} \, d\nu - \int_0^1 \psi \cdot q(\nu,t) \cdot x(\nu,t) \, d\nu.$$

The first-order condition with respect to any x gives the solution $\forall \nu$

$$x^{SP}(\nu,t) = \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}} L > x^{DE}(\nu,t) = \left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1}{\phi}} L, \qquad \text{since } 0 < \phi < 1.$$

It follows that $X^{SP}(t) = \psi L Q(t) \left(\frac{A}{\psi}\right)^{\frac{1}{\phi}}$. The planner uses more of each x than in the decentralized equilibrium because equilibrium prices are above marginal cost: $p = \psi q/(1-\phi) > \psi q$. Substituting x^{SP} in Y^{SP} gives:

$$Y^{SP}(t) = \frac{A}{1-\phi}\left(\frac{A}{\psi}\right)^{\frac{1-\phi}{\phi}} Q(t)L > Y^{DE}(t) = \frac{A}{1-\phi}\left(\frac{A(1-\phi)}{\psi}\right)^{\frac{1-\phi}{\phi}} Q(t)L$$

We can see that $Y^{SP} > Y^{DE}$. After simplifying, we get $X^{SP}(t) = (1-\phi)Y^{SP}(t)$, and it follows that $Y^{SP}(t) - X^{SP}(t) = \phi Y^{SP}(t)$. Using this the resource constraint reads

$$\phi Y^{SP}(t) = C^{SP}(t) + Z^{SP}(t).$$

Imposing symmetry, $z(\nu, t) = z(t)$, and BGP, where $\frac{\dot{Q}}{Q} = g = (\lambda - 1)z^{\star}$, yields:

$$\dot{Q}(t) = (\lambda - 1)z(t)Q(t)$$

Substituting $z(t)$, $Q(t)$, and $Z(t)$ from the budget constraint gives

$$\dot{Q}(t) = (\lambda-1)\eta\left[\phi Y^{SP}(t) - C(t)\right] = (\lambda-1)\eta\left[\phi\frac{A}{1-\phi}\left(\frac{A}{\psi}\right)^{\frac{1-\phi}{\phi}} Q(t)L - C(t)\right].$$

Hamiltonian of the problem is:

$$\mathcal{H} = e^{\rho t}\frac{C(t)^{1-\sigma}-1}{1-\sigma} + \mu(t)\left((\lambda-1)\eta\left[\phi\frac{A}{1-\phi}\left(\frac{A}{\psi}\right)^{\frac{1-\phi}{\phi}} Q(t)L - C(t)\right]\right).$$

Deriving the first-order conditions, and solving for the growth rate of consumption $C(t)$ yields:

$$g^{SP} = \frac{(\lambda-1)\eta\phi\frac{A}{1-\phi}\left(\frac{A}{\psi}\right)^{\frac{1-\phi}{\phi}} L - \rho}{\sigma}. \tag{13.A.28}$$

Note that g^{SP} is increasing in λ, η, L, and is decreasing in ρ, σ. In the decentralized equilibrium, we had $g^{\star} = g^{DE}$ given by (13.66). We see that $g^{SP} \neq g^{DE}$ for three reasons:

1. Business stealing: SP sees gain $(\lambda - 1)$ to innovation, whereas DE innovators see λ, a force for $g^{SP} < g^{DE}$.

2. Markup distortion: SP use intermediates more intensively $(1-\phi)^{\frac{1}{\phi}} > 1$ which increases the value of innovation: a force for $g^{SP} > g^{DE}$.

3. Knowledge externalities: SP sees increasing q lasting into future (innovators build on it), whereas profit decrease in future z. This is the $\left(\frac{1}{\lambda-1}\right)$ term in the denominator of g^{DE}. It's a force for $g^{SP} > g^{DE}$.

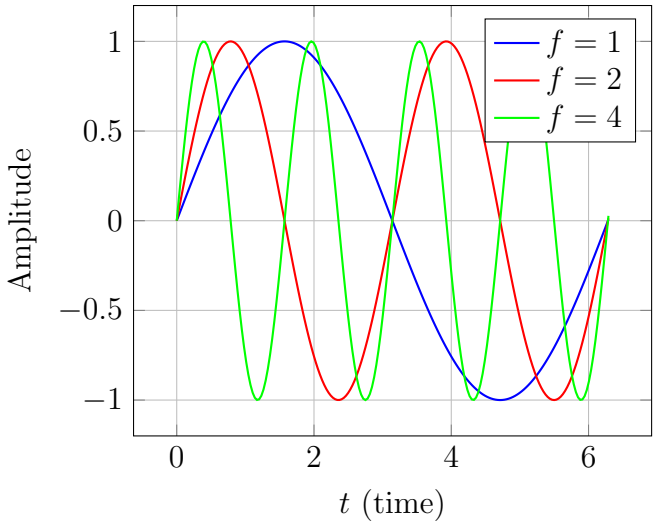


Figure 14.A.1: Sine waves with different frequencies

14.A Appendix to Chapter 14

Spectral decomposition and the band-pass filter

Spectrum of a process The spectrum of a time series process provides a decomposition of its variance across different frequencies. It is a fundamental tool in time series analysis, especially when studying cyclical patterns in the data. Given a stationary process Y_t, its spectrum, denoted as $s(\omega)$, is defined for each frequency ω in the interval $[0, \pi]$ as:

$$s(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-ik\omega}$$

where γ_k is the autocovariance function of the process at lag k (that is, $\gamma_k = Cov(Y_t, Y_{t-k})$ for all integers k). You may recognize that the spectrum is thus given by the Fourier transform of the autocovariance series $\{\gamma_k\}$.[33] Thus, we can recover the auto-covariance function by applying the inverse Fourier transform to the spectrum:

$$\int_{-\pi}^{\pi} s(\omega) e^{ik\omega} d\omega = \gamma_k$$

In order to derive this result we can substitute in the definitions of the spectrum:

$$\int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ij\omega} e^{ik\omega} d\omega = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j \int_{-\pi}^{\pi} e^{-ij\omega} e^{ik\omega} d\omega$$

Now, when $j = k$, $e^{-ik\omega} e^{ij\omega} = 1$, thus $\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik\omega} e^{ij\omega} d\omega = 1$ and we obtain γ_k. For $j \neq k$, $e^{-ik\omega} e^{ij\omega} = e^{i\omega(k-j)}$. Recall from Euler's Theorem that $e^{ix} = \cos(x) + i \sin(x)$. Thus, we can re-write the integral as $e^{k-j} \int_{-\pi}^{\pi} (\cos(\omega) + i \sin(\omega))\, d\omega$. That integral is equal to 0 for all $j \neq k$, since the integral of cosine and sine from $-\pi$ to π is zero. Thus, the entire expression collapses to γ_k.

[33] For a series $\{x_t\}$, the Fourier transform is $d_x(\omega) = \sum_{k=-\infty}^{\infty} x_t e^{-ik\omega}$.

Variance Decomposition using the Spectrum We can compute the variance by evaluating the above expression when $k = 0$, since the γ_0 is the variance of the process. The expression then simply becomes:

$$\text{Var}(Y_t) = \int_{-\pi}^{\pi} s(\omega)\, d\omega$$

The total variance of the process can be decomposed into the contributions from different frequencies. This is achieved by integrating the spectrum over different frequency intervals. For a given frequency band $[\omega_1, \omega_2]$, the contribution to the variance is:

$$\text{Var}_{\omega_1}^{\omega_2}(X_t) = \int_{\omega_1}^{\omega_2} f(\omega)\, d\omega$$

For a given frequency ω, the spectral density $s(\omega)$ gives the contribution to the variance of the process from cycles with frequency ω. Peaks in the spectral density indicate the presence of cycles at those frequencies. By examining the spectral density, one can identify the dominant frequencies or cycles present in the data. Peaks in the spectral density correspond to prominent cycles. The height of the peak indicates the strength or power of that particular cyclical component. For instance, a peak in the spectral density at a frequency corresponding to 40 quarters (or 10 years) would suggest a 10-year business cycle. One of the challenges of using the spectrum to try to identify cycles is that it requires stationary time series. Most of the aggregate data series (e.g. GDP, employment, investment) all exhibit strong trends. In order to perform the analysis requires removal of the trend, the process by which can change the underlying spectrum, making the interpretation of the results difficult.

Linear filters and band-pass filter A linear filter applies a linear transformation to the data, in the context of this chapter the aim of the filter is to separate out the cyclical vs trend components. Linear filters operate by combining present and past values of the time series. The filtered series Y_c can be written using a general lead-lag polynomial $C(L)$, where $Y_c = C(L)$. The HP filter is a linear filter where:

$$C(L) = \frac{\lambda(1-L)^2(1-L^{-1})^2}{1+\lambda(1-L)^2(1-L^{-1})^2}.$$

One useful tool for understanding and constructing linear filters is the *gain function*. The gain function of a linear filter quantifies how the filter amplifies or attenuates different frequency components of the time series. By analyzing how a linear filter affects different frequencies (as shown by its gain function), one can design a filter to achieve specific objectives, such as emphasizing or suppressing certain frequency ranges of the time series. Mathematically we can write the gain function of a particular linear filter as $G(\omega) = |C\left(e^{-i\omega}\right)| \equiv \sqrt{C\left(e^{-i\omega}\right) C\left(e^{i\omega}\right)}$. One can show that the spectrum of the filtered series (after applying the gain function) is given by:

$$s_{Y_c} = G(\omega)^2 s_Y(\omega).$$

Thus, if $|G(\omega)| > 1$ for some ω it amplifies that frequency, and if $|G(\omega)| < 1$ it attenuates it. To understand the gain function a bit better, let's consider a simple example. The first-differencing filter, $C(L) = (1 - L)$, is very commonly used for removing trends or to make

series stationary (as discussed earlier). It yields $Y_c = (1 - L)Y$. The gain function is thus given by $G(\omega) = |C(e^{-i\omega})| = \sqrt{(1 - e^{-i\omega})(1 - e^{i\omega})} = \sqrt{2(1 - \cos(\omega))}$. The spectrum is fully attenuated at $\omega = 0$, and very low for other frequencies close to zero. The gain function for the HP filter is a bit more complicated, but can be computed as

$$\frac{4\lambda(1 - \cos(\omega))^2}{1 + 4\lambda(1 - \cos(\omega))^2}$$

At very low frequencies or very long-term trends (when $\omega \approx 0$), $\cos(\omega) \approx 1$ implying that the gain function $G(\omega) \approx 0$. So long-term trends are attenuated and only cyclical frequencies are transmitted through. As λ gets large there is less smoothing, and the spectrum of the filtered series converges to that of the original time series.

Finally, the *band-pass filter* aims to extract cyclical components of a time series that lie within a specific frequency range. In particular, the gain function is 1 for the interval of frequencies selected, and 0 otherwise. Thus, the band-pass filter removes frequencies in a discrete way. The researcher selects to "window" of desired frequencies that should be kept (which can be of arbitrary size). As compared to the band-pass filter, both the HP filter and first difference filter remove frequencies smoothly and remove low frequencies more than high frequencies. A band-pass filter can be beneficial in studying business cycles if we thing that there is meaningful variation at higher than business-cycle frequencies, for example seasonal variation or a political process. In practice, however, most data are seasonally adjusted first, and then something like the HP filter is used to extract the cyclical component at business cycle frequencies.

Like the HP filter, the band-pass filter also has some limitations to be aware of. The choice of frequency range can influence the resulting decomposition. Different applications may require different frequency ranges. Like the HP filter, the band-pass filter may produce spurious results towards the beginning and end of the sample. Finally, like the HP filter, it doesn't rely on underlying economic theory.

Some examples of how the HP-filter and band-pass filter work are plotted in Figures (14.A.2) and (14.A.3). In both figures the time series $y_t = \sin(t) + 0.5\sin(5t)$ is plotted as the solid line. Notice that this is the mix of a low-frequency, high amplitude component ($\sin(t)$) and a high-frequency, lower-amplitude component ($0.5\sin(5t)$). The latter is higher frequency since it completes 5 cycles for every one cycle of the low-frequency component. In Figure (14.A.2), I applied the band-pass filter first to target the low-frequency component and then to the high-frequency one. Because of the periodic nature of this time series, we are able to recover the underlying components almost perfectly (except at the boundaries). In Figure (14.A.3) I instead apply the HP filter. The HP-filter is unable to perfectly separate the low- and high-frequency components as in the band-pass filter, but gets them approximately correct. You can still observe some low frequency variation in the cycle component.

Appendix Tables

Figure 14.A.2: Band-pass filter example

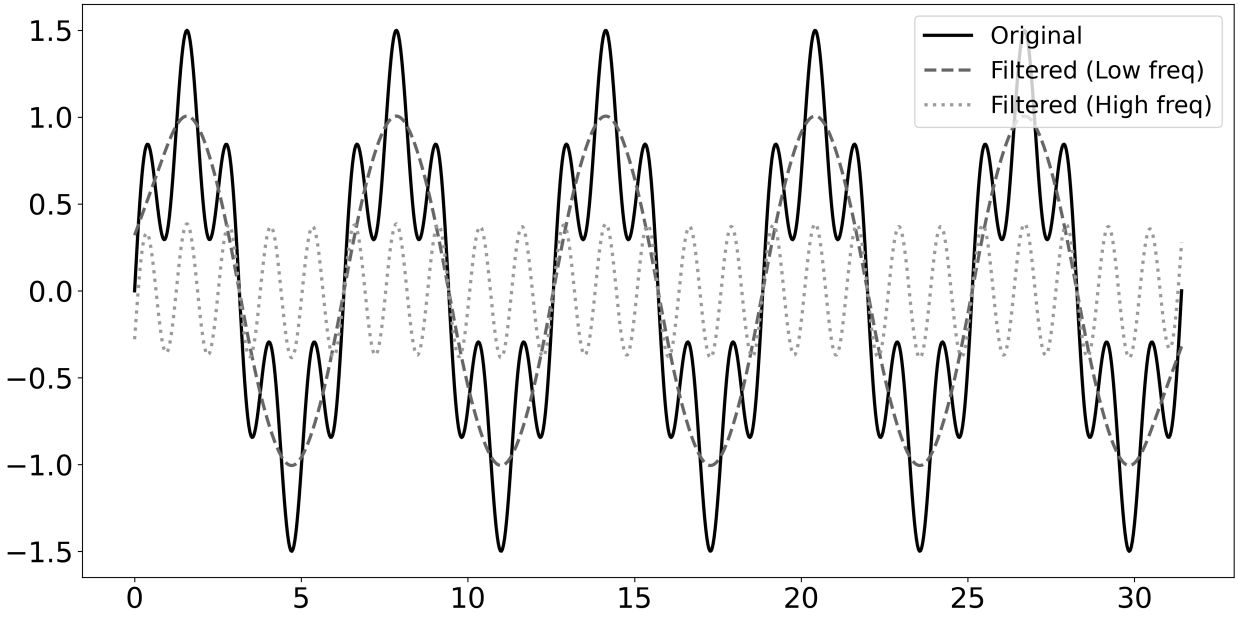


Figure 14.A.3: HP filter example

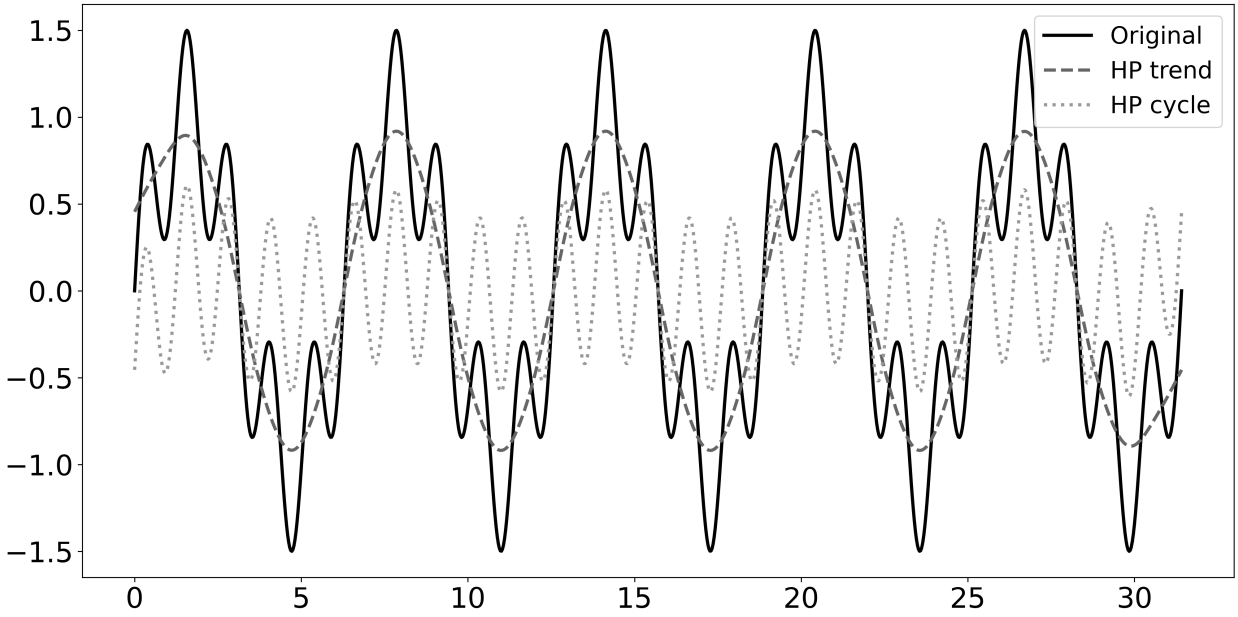


Table 14.A.1: Business cycle moments U.S. Data 1949-2022 (Baxter-King Filter)

Variable x	Standard Deviation	Relative Std to σ_y	Auto-correlation	Cross correlation of x with $y(t-1)$	$y(t)$	$y(t+1)$
Output (y)	0.014	1.000	0.910	0.910	1.000	0.910
Consumption	0.011	0.791	0.922	0.763	0.898	0.887
Gov. Consumption	0.022	1.551	0.961	0.140	0.098	0.007
Investment	0.061	4.263	0.909	0.787	0.892	0.839
Employment	0.013	0.926	0.935	0.897	0.798	0.579
Hours	0.018	1.253	0.929	0.893	0.869	0.707
Unemployment	0.129	8.999	0.916	-0.890	-0.862	-0.683
Labor Productivity	0.010	0.676	0.896	0.104	0.357	0.512
Wages	0.009	0.657	0.917	0.148	0.216	0.269
Price Level	0.008	0.554	0.964	-0.299	-0.382	-0.460
Fed Funds Rate	0.036	2.493	0.971	0.225	0.158	0.062

Notes: The particular band-pass filter is a linear approximation to a true band-pass filter developed by Baxter and King (1999). Per the suggestions of the paper, we focus on frequencies from 1.5-8 years.

Table 14.A.2: Business cycle moments U.S. Data 1949-2022 (First-difference Filter)

Variable x	Standard Deviation	Relative Std to σ_y	Auto-correlation	Cross correlation of x with $y(t-1)$	$y(t)$	$y(t+1)$
Output (y)	0.011	1.000	0.136	0.136	1.000	0.136
Consumption	0.011	0.962	-0.100	0.022	0.763	0.099
Gov. Consumption	0.017	1.459	0.567	0.190	0.158	0.055
Investment	0.046	4.013	0.165	0.145	0.765	0.170
Employment	0.010	0.902	0.116	0.275	0.766	-0.081
Hours	0.013	1.169	0.127	0.219	0.817	0.017
Unemployment	0.102	8.987	0.175	-0.316	-0.768	0.037
Labor Productivity	0.009	0.767	0.024	-0.137	0.404	0.223
Wages	0.009	0.833	-0.024	0.142	-0.143	0.217
Price Level	0.006	0.534	0.786	-0.036	-0.054	-0.114
Fed Funds Rate	0.009	0.753	0.259	0.223	0.284	0.056

15.A Appendix to Chapter 15

15.A.1 Data Appendix

The data on revenues, expenditures, and deficits is obtained from the Bureau of Economic Statistics (BEA), Table 3.1 "Government Current Receipts and Expenditures," which is part of the NIPA tables[34]. The series, expressed in current billion dollars, span the interval 1929-2021. They include Federal, State, and Local government budget measures (sometimes referred to as General Government or National Government statistics). In the plots, the series are expressed as percentages of "GDP," corresponding to *Gross Domestic Product* (line 1 of Table 1.5.5).

Figure 15.1: Data is obtained from the OECD Dataset: National Accounts at a Glance, and the variable corresponds to *Total expenditure of general government, percentage of GDP*.

Figure 15.2: "Revenues" correspond to *Total Receipts* (line 34 of Table 3.1) and "Expenditures" to *Total Expenditures* (line 37 of Table 3.1). We use total rather than current measures because these include public investment.

Figure 15.3: All series are obtained from Table 3.1 "Government Current Receipts and Expenditures," constructed by the BEA. "Income Taxes" corresponds to *Personal current taxes* (line 3), "Sales and Import Taxes" to *Taxes on production and imports* (line 4), "Corporate Taxes" to *Taxes on corporate income* (line 12), and "Social Ins. Taxes" to *Contributions for government social insurance* (line 7).

Figure 15.4 - Left Panel: "Total Deficit or Surplus," in Figure 15.4, is constructed as the difference between Revenues and Expenditures (defined above),

$$\text{Total Deficit}=\text{Revenues-Expenditures}.$$

"Net Interest" is the difference between *Interest and Miscellaneous Receipts* (line 11 of Table 3.1) and *Interest Payments* (line 27 of Table 3.1). The government simultaneously owns assets that yield interest and owes debt for which it has to pay interest. In the figure, we plot net interest payments. The "Primary Deficit" is defined as

$$\text{Primary Deficit}= \text{Total Deficit - Net Interest}.$$

Figure 15.4 - Right Panel: FRED provides debt series for Federal and State governments between 1946 and 2021. "Federal Debt" corresponds to *Federal Government; Debt Securities and Loans; Liability, Level* (or FGSDODNS), while "State and Local Debt" corresponds to *State and Local Governments; Debt Securities and Loans; Liability, Level* (or SLGSDODNS). Both series are obtained from the Flow of Funds tables constructed by the Board of Governors of the Federal Reserve Bank System. It is worth noticing that the Federal Debt series does

[34] See https://apps.bea.gov/iTable

not correspond exactly to the one provided by the White House historical series due to differences in accounting methods (i.e. which items are included and timing in which certain transactions are incorporated when computing the flow of funds). The series between 1916 and 1945 are obtained from the Survey of Current Business, September 1946 page 13, Table 5. They correspond to Net Public Debt, end of calendar year.

Figure 15.5 - Left Panel: All series are obtained from Table 3.1 (described above). "Govt Consumption (+ Investment)" is the sum of *Consumption expenditures* (line 20) and *Gross government investment* (line 39). "Transfers" is the sum of *Current transfer payments* (line 22) and *Subsidies* (line 30). "Interest Payments" are gross, obtained from line 27 (i.e. we are not including interest receipts). The sum of these is equal to Expenditures, defined above.

Expenditures= Govt Consumption (+ Investment) + Transfers + Interest Payments.

Figure 15.5 - Right Panel: All series are obtained from Table 3.16 "Government Current Expenditures by Function" constructed by the BEA. "Defense" corresponds to *National Defense* (line 7), "Healthcare" corresponds to *Health* (line 28), and "Education" is obtained directly form line 30. "Income Security" is obtained from line 36. It includes *Disability* (line 37), *Welfare and social services* (line 39), *Unemployment* (line 40), *Retirement* (line 38) and other income insurance programs (line 41). "Other" is constructed as the sum of *General public service* (line 2), *Public order and safety* (line 8), *Economic affairs* (line 13), *Housing and community services* (line 27), *Recreation and culture* (line 29), minus *Interest payments* (line 5).

15.A.2 Tax reform with wealth effects

In this section, we re-compute the tax reform from Section 15.3.3, but assuming that utility takes the form

$$u(c,\ell) = \ln c - \frac{\ell^{1+\frac{1}{\phi}}}{1+\frac{1}{\phi}}.$$

The first order condition with respect to labor implies

$$\ell^{1/\phi} = \left(\frac{1-\tau_t^l}{1+\tau_t^c}\right) w \frac{1}{c_w} \quad \text{with} \quad c_w = \left(\frac{1-\tau_t^l}{1+\tau_t^c}\right) w.$$

The labor supply is *independent* of taxes in this case, $\ell = 1$. This happens because the substitution effect, that would make ℓ decline when after-tax labor income goes down is exactly offset by the income effect, caused by a decline in c_w that results in lower after-tax income.

The main difference between Figure 24.4 and Figure 15.A.1 is that now labor supply remains constant when labor taxes are increased. As a result, we do not observe a decline in output, which allows the government to have higher G (recall that the exercise is constructed such that $G/Y = 0.2$ throughout the simulation). In the long-run, because labor does not go down, there is higher GDP and aggregate consumption is slightly higher. The tax reform is more effective in this scenario because the costs of replacing capital taxes with labor taxes are smaller when wealth effects are present.

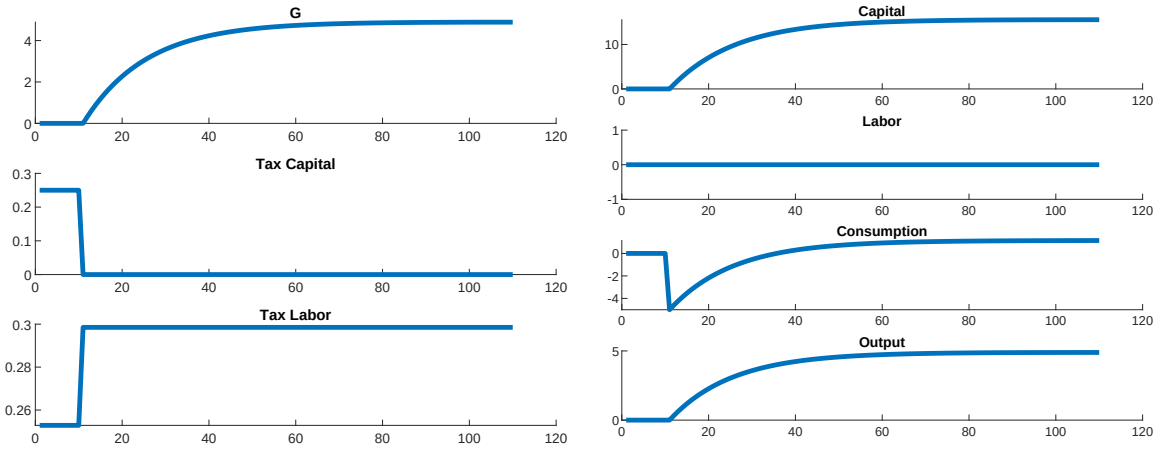


Figure 15.A.1: Eliminating capital income taxes (with wealth effects)

18.A Appendix to Chapter 18

18.A.1 Derivation of the New Keynesian Phillips curve

The envelope condition of the intermediate goods producer's Bellman equation is

$$V_p(p,\mathcal{S}) = u'(C(\mathcal{S}))Y(\mathcal{S})\left[p^{-\varepsilon} - \varepsilon p^{-\varepsilon-1}\left(p - \frac{w(\mathcal{S})}{A(\mathcal{S})}\right)\right] + \beta\mathbb{E}\left[\theta V_p\left(\frac{p}{1+\pi(\mathcal{S}')}, \mathcal{S}'\right)\frac{1}{1+\pi(\mathcal{S}')}\right]$$

and the first-order condition of the price-setting problem is

$$V_p(p,\mathcal{S}) = 0.$$

V_p gives the benefit to the firm of having a higher price. Notice that the first term on the right-hand side of the envelope condition is the marginal increase in profit this period from setting a higher price (multiplying by $u'(C)$ values this change in profit in terms of utility). The envelope condition then has the form of an expected discounted sum of marginal changes in profits today and, if prices are sticky ($\theta > 0$), in the future.

Combining the envelope condition and first order condition, the solution to the price-setting problem at date t, p_t^R, must satisfy

$$0 = \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} u'(C_\tau) Y_\tau \left[{p_{t,\tau}^R}^{-\varepsilon} - \varepsilon {p_{t,\tau}^R}^{-\varepsilon-1}\left(p_{t,\tau}^R - \frac{w_\tau}{A_\tau}\right)\right],$$

where $p_{t,\tau}^R \equiv P_t^R/P_\tau = p_t^R/\prod_{s=t+1}^{\tau}(1+\pi_s)$, is the relative price at date τ of a firm that last updated its price at date t. Rearranging we obtain

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} u'(C_\tau) Y_\tau {p_{t,\tau}^R}^{-\varepsilon} = \frac{\varepsilon}{\varepsilon-1}\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} u'(C_\tau) Y_\tau {p_{t,\tau}^R}^{-\varepsilon-1}\frac{w_\tau}{A_\tau}.$$

We now log-linearize both sides of this equation around a zero-inflation steady state in which $p_{t,\tau}^R = 1$ for all τ to obtain

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t} \hat{p}_{t,\tau}^R = \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t}\left(\hat{w}_\tau - \hat{A}_\tau\right), \tag{18.A.1}$$

where hats denote log deviations from steady state. From the definition of $p_{t,\tau}^R$ we have[35]

$$\hat{p}_{t,\tau}^R = \hat{P}_t^R - \hat{P}_\tau = \hat{P}_t^R - \hat{P}_t - (\hat{P}_\tau - \hat{P}_t) = \hat{p}_t^R - (\hat{P}_\tau - \hat{P}_t).$$

Eq. (18.A.1) then becomes

$$\begin{aligned}
\hat{p}_t^R &= (1-\beta\theta)\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta)^{\tau-t}\left[\hat{w}_\tau - \hat{A}_\tau + \hat{P}_\tau - \hat{P}_t\right], \\
&= (1-\beta\theta)\left(\hat{w}_t - \hat{A}_t\right) + (1-\beta\theta)\mathbb{E}_t \sum_{\tau=t+1}^{\infty} (\beta\theta)^{\tau-t}\left[\hat{w}_\tau - \hat{A}_\tau + \hat{P}_\tau - \hat{P}_t\right] \\
&= (1-\beta\theta)\left(\hat{w}_t - \hat{A}_t\right) + \beta\theta\mathbb{E}_t\left(\hat{p}_{t+1}^R + \hat{P}_{t+1} - \hat{P}_t\right).
\end{aligned} \tag{18.A.2}$$

[35] Here we use $\widehat{(1+\pi_s)} = \pi_s$.

Note that $\pi_{t+1} \approx \log(1+\pi_{t+1}) = \hat{P}_{t+1} - \hat{P}_t$. Log-linearizing eq. (18.10) yields

$$\pi_t = \frac{1-\theta}{\theta}\hat{p}_t^R \tag{18.A.3}$$

and combining this with eq. (18.A.2) yields

$$\pi_t = \frac{(1-\theta)(1-\beta\theta)}{\theta}\left(\hat{w}_t - \hat{A}_t\right) + \beta\mathbb{E}_t\pi_{t+1}. \tag{18.A.4}$$

This equation has the form of a New Keynesian Phillips curve in which the term $\hat{w}_t - \hat{A}_t$ is the (log-linearized) real marginal cost of producing goods. The last step is to express this marginal cost in terms of the output gap. We do that using the production function, the aggregate resource constraint, and the household's labor supply condition.

Log-linearizing the aggregate production function, eq. (18.11), yields

$$\hat{Y}_t = \hat{A}_t - \hat{D}_t + \hat{L}_t$$

and log-linearizing eq. (18.12) around the steady state values $\bar{D} = \bar{p}^R = 1 + \bar{\pi} = 1$ yields

$$\hat{D}_t = -\varepsilon(1-\theta)\hat{p}_t^R + \theta\varepsilon\pi_t + \theta\hat{D}_{t-1}$$

and using (18.A.3) this simplifies to $\hat{D}_t = \theta\hat{D}_{t-1}$. As there is no price dispersion in steady state, $\hat{D}_t = \theta\hat{D}_{t-1}$ implies that $\hat{D}_t = 0$ for all t. That is, price dispersion does not affect the first-order approximation to the dynamics of the economy.

Log-linearizing the labor supply condition,(18.6), we obtain

$$-\sigma\hat{C}_t + \hat{w}_t = \psi\hat{L}_t.$$

Now using the resource constraint $Y_t = C_t$ and the log-linearized production function we have

$$\hat{w}_t = (\sigma+\psi)\hat{Y}_t - \psi A_t. \tag{18.A.5}$$

Log-linearizing (18.13) we obtain

$$Y_t^n = \frac{1+\psi}{\psi+\sigma}\hat{A}_t. \tag{18.A.6}$$

Finally, we can combine (18.A.5) and (18.A.6) to obtain

$$\hat{w}_t - \hat{A}_t = (\sigma+\psi)(\hat{Y}_t - \hat{Y}_t^n).$$

We then plug this into (18.A.4) to obtain equation (18.16).

18.A.2 Taylor Principle

This explanations draws on Bullard and Mitra (2002). We are interested in the stability of the dynamic system defined by the IS curve (18.15), the Phillips curve (18.16), and the interest rate rule (18.17). Here we substitute out for the nominal interest rate using the interest rate rule. We will study a deterministic economy, which is without loss of generality given the certainty equivalent property of first-order accurate economies. The system can then be expressed as

$$\begin{aligned}\beta\sigma\hat{Y}_{t+1} &= \beta\sigma\hat{Y}_t + \beta\phi_\pi\pi_t + \beta\phi_x\hat{Y}_t - \pi_t + \kappa\hat{Y}_t \\ \beta\pi_{t+1} &= \pi_t - \kappa\hat{Y}_t,\end{aligned}$$

where the first equation is the IS curve with π_{t+1} substituted out using the Phillips curve. For simplicity, we are assuming $\pi^* = 0$. We can then write this system as

$$\begin{pmatrix}\beta\sigma & 0\\ 0 & \beta\end{pmatrix}\begin{pmatrix}\hat{Y}_{t+1}\\ \pi_{t+1}\end{pmatrix} = \begin{pmatrix}\beta\sigma + \beta\phi_x + \kappa & \beta\phi_\pi - 1\\ -\kappa & 1\end{pmatrix}\begin{pmatrix}\hat{Y}_t\\ \pi_t\end{pmatrix}.$$

Inverting the coefficient matrix on the right-hand side we have

$$\frac{1}{\sigma + \phi_x + \kappa\phi_\pi}\begin{pmatrix}\sigma & 1 - \beta\phi_\pi\\ \kappa\sigma & \beta\sigma + \beta\phi_x + \kappa\end{pmatrix}\begin{pmatrix}\hat{Y}_{t+1}\\ \pi_{t+1}\end{pmatrix} = \begin{pmatrix}\hat{Y}_t\\ \pi_t\end{pmatrix}.$$

This model has no state variables and two forward-looking variables. For a unique equilibrium, we need both eigenvalues of the coefficient matrix that multiplies $(\hat{Y}_{t+1}, \pi_{t+1})$ to be inside the unit circle. The characteristic polynomial of this matrix is $p(\lambda) \equiv \lambda^2 + a_1\lambda + a_0$ where

$$\begin{aligned}a_1 &= -\left(\frac{\sigma + \beta\sigma + \beta\phi_x + \kappa}{\sigma + \phi_x + \kappa\phi_\pi}\right)\\ a_0 &= \frac{\beta\sigma}{\sigma + \phi_x + \kappa\phi_\pi}.\end{aligned}$$

When are the roots of this polynomial inside the unit circle? We can answer this using the Jury Stability Criterion,[36] which in this specific case (a quadratic polynomial with a unit coefficient on λ^2) requires that $|a_0| < 1$ and $|a_1| < 1 + a_0$. The former condition can be expressed as $-(1-\beta)\sigma < \phi_x + \kappa\phi_\pi$, which holds as we assume $\beta \in [0, 1]$, all the parameters are weakly positive and σ is strictly positive. Turning to the condition $|a_1| < 1 + a_0$, this can be expressed as condition (18.18).

18.A.3 A model with nominal wage and price rigidities

The representative household has the same preferences as in the sticky-price model except we now denote hours worked by N_t while L_t will refer to effective labor as we explain below. So preferences are

$$U_0 = \mathbb{E}_0\sum_{t=0}^{\infty}\beta^t\left[\frac{C_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\psi}}{1+\psi}\right]. \tag{18.A.7}$$

[36] See the discussion of the related Schur-Cohn Criterion in on p. 27 of LaSalle (1986).

This labor is supplied to a continuum of labor unions that differentiate it. Let $n_{i,t}$ be the amount of labor used by union $i \in [0,1]$.

The technology for producing goods is the same as in eqs. (18.2) and (18.3). In particular, $\ell_{j,t}$ remains the amount of labor used in producing good j. The labor used in producing the final good is a composite of the various types supplied by the unions. Specifically, the total effective labor supply is

$$L_t = \left(\int_0^1 n_{i,t}^{\frac{\varepsilon-1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{18.A.8}$$

and market clearing requires $\int_0^1 \ell_{j,t} dj = L_t$. Here we have assumed the elasticity of substitution between labor varieties is the same as that between intermediate goods. This assumption is not necessary and we make it only for the sake of highlighting the similarity between the sticky-wage and sticky-price models.

Let $W_{i,t}$ be the nominal wage of type-i labor—that is, the labor supplied by union i. The cost-minimization problem of intermediate firm j implies firm j's demand for type-i labor is $(W_{i,t}/W_t)^{-\varepsilon}\ell_{j,t}$, where W_t is the cost of producing one unit of labor aggregate. This aggregate wage index is

$$W_t = \left(\int_0^1 W_{i,t}^{1-\varepsilon} di \right)^{1/(1-\varepsilon)}, \tag{18.A.9}$$

Total labor income across all unions is

$$\int_0^1 \int_0^1 W_{i,t}(W_{i,t}/W_t)^{-\varepsilon}\ell_{j,t} dj di = \int_0^1 W_{i,t}(W_{i,t}/W_t)^{-\varepsilon} di L_t = W_t L_t.$$

Effective labor supply, L_t, is related to hours worked as follows. Total hours worked must equal the total usage by the unions $N_t = \int_0^1 n_{i,t} di$. In turn, total labor usage by union i must equal the total supplied to each firm

$$n_{i,t} = \int_0^1 (W_{it}/W_t)^{-\varepsilon}\ell_{jt} dj = (W_{it}/W_t)^{-\varepsilon} L_t.$$

Putting the two together we have

$$N_t = \underbrace{\int_0^1 (W_{it}/W_t)^{-\varepsilon} di}_{\equiv D_t^W} L_t.$$

The term $D_t^W \geq$ reflects wage dispersion, which results in hours worked exceeding the effective labor supply as work effort is inefficiently allocated across types of labor.

We introduce a nominal rigidity for wages that is analogous to the one we assumed for prices. Each period, union i is able to update the wage for i-type labor with probability $1-\theta_w \in [0,1]$. As in the intermediate goods firm's problem, the union must supply whatever quantity of labor is demanded at the prevailing price. By participating in the union, the

household agrees to work the amount of hours the union needs in exchange for the wage the union sets.

The household takes aggregate labor income and hours worked as given as determined by the labor union. Their Euler equation is unchanged from the sticky-price model. By the envelope theorem, the marginal value of funds at date t is $u'(C_t)$ and the marginal disutility of labor supply is N_t^{ψ}.

We now turn to the union's wage-setting problem, which is quite similar to the firm's price-setting problem. When raising the wage, the firm raises labor income but this comes at a disutility cost. As the wage is sticky, the union makes a forward-looking choice. The objective of the union that sets its wage at date t is

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} \left[u'(C_\tau) \frac{W_t^R}{P_\tau} - N_\tau^{\psi} \right] \left(\frac{W_t^R}{W_\tau} \right)^{-\varepsilon} L_\tau,$$

where W_t^R is the wage chosen at date t. In this objective, we assume that real labor income is valued at $\beta^{\tau-t}u'(C_\tau)$. Each union takes this marginal utility as given because each union contributes an infinitesimal part of the total income of the household. Similarly, the union takes the disutility of $\beta^{\tau-t}N_\tau^{\psi}$ of work effort as given. The first order condition w.r.t. W_t^R is

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} u'(C_\tau) w_\tau \left(w_{t,\tau}^R \right)^{-(\varepsilon-1)} L_\tau = \frac{\varepsilon}{\varepsilon-1} \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} N_\tau^{\psi} \left(w_{t,\tau}^R \right)^{-\varepsilon} L_\tau,$$

where $w_t \equiv W_t/P_t$ and $w_{t,\tau}^R \equiv W_t^R/W_\tau$. Log-linearizing around a zeros inflation steady state we obtain

$$\mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} \hat{w}_{t,\tau}^R = \mathbb{E}_t \sum_{\tau=t}^{\infty} (\beta\theta^w)^{\tau-t} \left(\sigma\hat{C}_\tau + \psi\hat{N}_\tau - \hat{w}_\tau \right). \tag{18.A.10}$$

Notice this is analogous to (18.A.1) with the main difference being that log marginal cost, $\hat{w}_\tau - \hat{A}_\tau$, is replaced with the difference between the log marginal rate of substitution between consumption and leisure, $\sigma\hat{C}_\tau + \psi\hat{N}_\tau$ and the log real wage, $\hat{w}_\tau$. The rest of the derivation of the wage Phillips curve follows steps similar to those for the price Phillips curve so we will omit many of the details here. Defining $\hat{w}_t^R \equiv \hat{W}_t^R - \hat{W}_t$ we have

$$\hat{w}_{t,\tau}^R = \hat{W}_t^R - \hat{W}_t - (\hat{W}_\tau - \hat{W}_t) = \hat{w}_t^R - (\hat{W}_\tau - \hat{W}_t)$$

and using this, (18.A.10) becomes

$$w_t^R = (1 - \beta\theta^w) \left(\sigma\hat{C}_t + \psi\hat{N}_t - \hat{w}_t \right) + \beta\theta^w \mathbb{E}_t \left(\hat{w}_{t+1}^R + \pi_{t+1}^w \right).$$

As in the sticky-price model we have

$$\pi_t^w \equiv \frac{W_t}{W_{t-1}} - 1 \approx \frac{1-\theta^w}{\theta^w} \hat{w}_t^R$$

so we have

$$\pi_t^w = \xi^w \left(\sigma\hat{C}_t + \psi\hat{N}_t - \hat{w}_t \right) + \beta\mathbb{E}_t \left[\pi_{t+1}^w \right],$$

where

$$\xi^w \equiv \frac{(1-\beta\theta^w)(1-\theta^w)}{\theta^w}.$$

The log-linearized aggregate production function is now $\hat{Y}_t = \hat{A}_t + \hat{L}_t = \hat{A}_t + \hat{N}_t$ and the log-linearized aggregate resource constraint remains $\hat{Y}_t = \hat{C}_t$. These lead to

$$\sigma\hat{C}_t + \psi\hat{N}_t = (\sigma+\psi)\hat{Y}_t - \psi\hat{A}_t.$$

As the flexible-price economy is unaffected relative to the sticky-price model, we still have $\hat{Y}_t^n = (1+\psi)/(\sigma+\psi)\hat{A}_t$. The wage Phillips curve then becomes

$$\pi_t^w = \kappa^w x_t + \xi^w\hat{A}_t - \xi^w\hat{w}_t + \beta\mathbb{E}_t\left[\pi_{t+1}^w\right],$$

where $\kappa^w \equiv \xi^w(\sigma+\psi)$.

Turning to the price Phillips curve, the derivation is the same as in Appendix 18.A.1 up to eq. (18.A.4). We therefore have

$$\pi_t = \xi^p\hat{w}_t - \xi^p\hat{A}_t + \beta\mathbb{E}_t\left[\pi_{t+1}\right],$$

where $\xi^p \equiv (1-\beta\theta)(1-\theta)/\theta$.

19.A Appendix to Chapter 19

19.A.1 Derivation of firm's first order conditions

Denote by λ the Lagrange multiplier associated with the budget constraint and $\mu\lambda$ the Lagrange multiplier associated with the borrowing constraint. The first order conditions for Problem (19.3) are

$$\begin{aligned}
&1 = \lambda\varphi_d(d),\\
&F_l(z,k,l) = w,\\
&\mathbb{E}m'\Omega_{k'}(S';k',b') - \lambda p + \mu\lambda\xi p = 0,\\
&\mathbb{E}m'\Omega_{b'}(S';k',b') + \frac{\lambda}{\tilde{R}} - \frac{\mu\lambda}{\tilde{R}} = 0,
\end{aligned}$$

where a subscript on a function denotes the derivative with respect to the particular variable. As observed earlier, the firm is atomistic and, therefore, the impact of its policies on the discount factor m' and, more generally, on the next period aggregate states, is negligible. Therefore, in the derivation of the first order conditions we took m' as given.

Next we differentiate the value function with respect to k' and b', which returns the envelope conditions,

$$\begin{aligned}
\Omega_k(S;k,b) &= \lambda\Big[(1-\delta)p + F_k(z,k,l)\Big],\\
\Omega_b(S;k,b) &= -\lambda.
\end{aligned}$$

We use envelope conditions, updated by one period, in the first order conditions to eliminate the next period derivatives of the value function $\Omega_{k'}(S';k',b')$ and $\Omega_{b'}(S';k',b')$. Furthermore, eliminating the multiplier λ using the first order condition for d, we can rewrite the first order conditions of the firm as in (19.4), (19.5) and (19.6).

19.A.2 Derivation of household's first order conditions

We derive the first order conditions for Problem (19.7) with respect consumption, c, labor, ℓ, new holding of bonds, b', and new holding of firms' shares, a'. Denoting by γ the Lagrange multiplier associated with the budget constraint, the first order conditions are

$$\begin{aligned}
&u_c(c,\ell) = \gamma,\\
&u_\ell(c,\ell) + \gamma w = 0,\\
&\beta\mathbb{E}V_b(S';a',b') - \gamma\frac{1}{R} = 0,\\
&\beta\mathbb{E}V_a(S';a',b') - \gamma q = 0.
\end{aligned}$$

The envelope conditions are derived by differentiating the value function with respect to b' and a',

$$\begin{aligned}
V_b(S;a,b) &= \gamma,\\
V_a(S;a,b) &= \gamma(d+q).
\end{aligned}$$

We use the envelope conditions, updated by one period, to eliminate $V_b(S';a',b')$ and $V_a(S';a',b')$ in the first order conditions. We then eliminate the multiplier γ using the first order condition for c to obtain equations (19.8), (19.9) and (19.10).

20.A Appendix to Chapter 20

20.A.1 Detailed derivation of the Generalized Nash Bargaining solution

Consider the problem

$$\max_w (\tilde{W}(w,z) - U(z))^\gamma (\tilde{J}(w,z) - V(z))^{1-\gamma},$$

where

$$\tilde{W}(w,z) = w + \beta \mathbb{E}[(1-\sigma)W(z') + \sigma U(z')],$$
$$\tilde{J}(w,z) = z - w + \beta \mathbb{E}[(1-\sigma)J(z') + \sigma V(z')],$$

and $U(z)$ and $V(z)$ are defined in (20.12) and (20.8). Note that these definitions imply $\partial \tilde{W}(w,z)/\partial w = 1$ and $\partial \tilde{J}(w,z)/\partial w = -1$.

The first-order condition for the maximization problem above is

$$\frac{\partial \tilde{W}(w,z)}{\partial w} \gamma (\tilde{W}(w,z) - U(z))^{\gamma-1} (\tilde{J}(w,z) - V(z))^{1-\gamma} = \frac{\partial \tilde{J}(w,z)}{\partial w} (1-\gamma)(\tilde{W}(w,z) - U(z))^{\gamma} (\tilde{J}(w,z) - V(z))^{-\gamma}.$$

Using $\partial \tilde{W}(w,z)/\partial w = 1$ and $\partial \tilde{J}(w,z)/\partial w = -1$ and reorganizing, we obtain (20.13).

20.A.2 Analysis of wages in the basic DMP model in Section 20.4

From the six equations (20.7), (20.8), (20.9), (20.11), (20.12), and (20.13), we can derive the equilibrium wages. In the DMP model, the worker's marginal product z and the opportunity cost of working b are different, and the surplus $z - b$ is divided between firms and workers. A simplistic wage rule could then be giving the workers $b + \gamma(z-b)$, where $\gamma \in (0,1)$ is the parameter that governs the workers' bargaining power. The actual solution in the basic DMP model is more complex because the Nash bargaining solution is over the present value of surpluses. Even though the ultimate source of the surplus is $z - b$, how it is divided between $w - b$ (workers) and $z - w$ (firms) is affected by the future possibilities.

Using (20.9) on (20.7) and multiplying γ,

$$\gamma J(z) = \gamma(z - w) + \beta \mathbb{E}[\gamma(1-\sigma)J(z')]$$

holds, and subtracting (20.12) from (20.11) and multiplying $(1-\gamma)$,

$$(1-\gamma)(W(z) - U(z)) = (1-\gamma)(w-b) + \beta \mathbb{E}[(1-\gamma)(1-\sigma-\lambda_w(\theta))(W(z') - U(z'))]$$

holds. Combining these two and using (20.13),

$$w = b + \gamma(z-b) + \beta \mathbb{E}[(1-\gamma)\lambda_w(\theta)(W(z') - U(z'))].$$

Thus, the wage is equal to the static share of the surplus $b + \gamma(z-b)$ plus the term that involves $\lambda_w(\theta)$. The latter term arises because the worker loses the opportunity of searching for a new job by being matched, and the worker is compensated for this loss. Using (20.13) and (20.10), the expression for the wage can alternatively be rewritten as

$$w = b + \gamma(z-b) + \gamma\theta\kappa.$$

20.A.3 Method of log-linearization

One method of analyzing macroeconomic models (especially business cycle models) is to first approximate the model by a system of linear equations. One popular method is to log-linearize the equilibrium condition. The advantage of the log-linearization is that the outcome is in terms of percentage deviations and therefore easy to interpret.

There are many methods of log-linearization. One method of carrying out the log-linearization of equations is following the next steps:

Step 1. Rewrite variable X_t as $X_t = \bar{X}e^{x_t}$, where $x_t \equiv \log(X_t/\bar{X})$ and $\bar{X}$ is the steady-state value of X. x_t is the percent deviation of X_t from its steady-state level.

Step 2. After simplifying, use the approximation $\bar{X}e^{x_t} \approx \bar{X}(1 + x_t)$ to (log-)linearize the equation.

Step 3. Further simplify the equation using the steady-state relationship. Note that expected value for the future can easily be dealt with because the expectation operator is linear.

As an example, consider the evolution of capital stock in the neoclassical growth model: $K_{t+1} = K_t^{\alpha}N_t^{1-\alpha} + (1-\delta)K_t - C_t$. The steady-state relationship is $\bar{K} = \bar{K}^{\alpha}\bar{N}^{1-\alpha} + (1-\delta)\bar{K} - \bar{C}$. The first step is to rewrite the equation as $\bar{K}e^{k_{t+1}} = (\bar{K}e^{k_t})^{\alpha}(\bar{N}e^{n_t})^{1-\alpha} + (1-\delta)\bar{K}e^{k_t} - \bar{C}e^{c_t}$. Simplifying, $\bar{K}e^{k_{t+1}} = \bar{K}^{\alpha}\bar{N}^{1-\alpha}e^{\alpha k_t + (1-\alpha)n_t} + (1-\delta)\bar{K}e^{k_t} - \bar{C}e^{c_t}$. Following the step 2 yields, $\bar{K}(1 + k_{t+1}) = \bar{K}^{\alpha}\bar{N}^{1-\alpha}(1 + \alpha k_t + (1-\alpha)n_t) + (1-\delta)\bar{K}(1 + k_t) - \bar{C}(1 + c_t)$. In step 3, we obtain $\bar{K}k_{t+1} = \bar{K}^{\alpha}\bar{N}^{1-\alpha}(\alpha k_t + (1-\alpha)n_t) + (1-\delta)\bar{K}k_t - \bar{C}c_t$.

As we saw above, if the equilibrium conditions are power functions, log-linearization is fairly straightforward. When they involve more complex functions, it may be useful to first Taylor approximate the original function $f(X_t)$ by

$$f(X_t) \approx f(\bar{X}) + f'(\bar{X})(X_t - \bar{X})$$

and then apply the above method. Because the approximated function is linear in X_t, it is straightforward to apply the above method.

20.A.4 Log-linarization of Section 20.7.2

This Appendix derives the log-linearized system in Section 20.7.2. With the Cobb-Douglas matching function, we can rewrite (20.31) as

$$\frac{\kappa}{(1-\gamma)\beta\chi}\theta_t^{\eta} = \beta\mathbb{E}\left[z_{t+1} - c(z_{t+1}) - b + \frac{\kappa\theta_{t+1}^{\eta}}{(1-\gamma)\chi} - \frac{\kappa\sigma(z_{t+1})\theta_{t+1}^{\eta}}{(1-\gamma)\chi} - \frac{\gamma\kappa\theta_{t+1}}{1-\gamma}\right].$$

In (20.33) and (20.34), we saw that $\hat{\sigma}(z_t)$ and $\hat{c}(z_t)$ can be approximated as

$$\hat{\sigma}(z_t) = \mathcal{E}\hat{z}_t$$

and

$$\hat{c}(z_t) = \mathcal{F}\hat{z}_t.$$

Then, the above equation can further be rewritten as

$$\frac{\kappa}{(1-\gamma)\beta\chi}\bar{\theta}^{\eta}e^{\eta\hat{\theta}_t} = \beta\mathbb{E}\left[\bar{z}e^{\hat{z}_{t+1}} - \bar{c}e^{\mathcal{F}\hat{z}_{t+1}} - b + \frac{\kappa\bar{\theta}^{\eta}e^{\eta\hat{\theta}_{t+1}}}{(1-\gamma)\chi} - \frac{\kappa\bar{\sigma}e^{\mathcal{E}\hat{z}_{t+1}}\bar{\theta}^{\eta}e^{\eta\hat{\theta}_{t+1}}}{(1-\gamma)\chi} - \frac{\gamma\kappa\bar{\theta}e^{\hat{\theta}_{t+1}}}{1-\gamma}\right].$$

Using the approximation $1+x \approx e^x$ and the steady-state relationship,

$$\frac{\kappa}{(1-\gamma)\beta\chi}\bar{\theta}^{\eta}\eta\hat{\theta}_t = \beta\mathbb{E}\left[\bar{z}\hat{z}_{t+1} - \bar{c}\mathcal{F}\hat{z}_{t+1} + \frac{\kappa\bar{\theta}^{\eta}\eta\hat{\theta}_{t+1}}{(1-\gamma)\chi} - \frac{\kappa\bar{\sigma}\bar{\theta}^{\eta}(\mathcal{E}\hat{z}_{t+1}+\eta\hat{\theta}_{t+1})}{(1-\gamma)\chi} - \frac{\gamma\kappa\bar{\theta}\hat{\theta}_{t+1}}{1-\gamma}\right].$$

With the same procedure as in the exogenous separation case, we obtain

$$\hat{\theta}_t = \mathcal{G}\hat{z}_t.$$

where

$$\mathcal{G} = (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\left(\bar{z} - \bar{c}\mathcal{F} - \frac{\kappa\bar{\sigma}\bar{\theta}^{\eta}\mathcal{E}}{(1-\gamma)\chi}\right). \tag{20.A.1}$$

We can solve for $\mathcal{G}$ from (20.A.1) after plugging in $\mathcal{E}$ and $\mathcal{F}$ from (20.33) and (20.34).

The solution is

$$\mathcal{G} = \frac{\Theta}{\Gamma},$$

where

$$\Theta \equiv (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{z}$$

and

$$\Gamma \equiv 1 + (1-\gamma)\left[\frac{\kappa\bar{\theta}^{\eta}\eta}{\chi}\left(\frac{1}{\rho\beta} - (1-\bar{\sigma})\right) + \kappa\gamma\bar{\theta}\right]^{-1}\bar{c}\frac{\xi\eta}{\xi+1}\left(1 - \frac{1}{1-\gamma}\right).$$

20.A.5 Derivation of equation (20.38)

First, modify the model so that the families can issue and sell/buy Arrow securities (contingency claims) for the next period state. Because the Arrow securities' net supply is zero and the families are identical (similar to the Lucas "tree" model), the existence of the Arrow securities does not affect the equilibrium allocation. Consider the problem

$$\max_{\{c_t, k_{t+1}, a_{t+1}(z_{t+1})\}_{t=0}^{\infty}} \mathbb{E}_0\left[\sum_{t=0}^{\infty}\mathbf{U}(c_t)\right],$$

$$c_t + k_{t+1} + \int Q_t(z_{t+1})a_{t+1}(z_{t+1})dz_{t+1} = (1+r_t-\delta)k_t + (1-u_t)w_t + u_t b + d_t + a_t(z_t),$$

where $Q_t(z_{t+1})$ is the price of an Arrow security that is issued (and traded) in period t and pays one unit of consumption goods if the next period state turns out to be z_{t+1}. $a_{t+1}(z_{t+1})$ is the quantity the family purchases at period t. Note that because the sum of a_t is zero and because the families are homogeneous, the equilibrium values of a_t are going to be zero. Therefore, existence of the Arrow securities do not alter the equilibrium allocation of goods.

With the recursive formulation,

$$\mathbf{V}(k, a(z), X) = \max_{c,k',\{a'(z')\}} \mathbf{U}(c) + \beta \int \mathbf{V}(k', a(z'), X') f(z'|z) dz'$$

subject to

$$c + k' + \int Q(z', X) a'(z') dz' = (1 + r(X) - \delta)k + (1 - u)w(X) + ub + d(X) + a(z),$$

$$K' = \Omega(X),$$

and

$$u' = (1 - \lambda_w(\theta(X))) + \sigma(1 - u).$$

Letting the Lagrange multiplier on the budget constraint when the stochastic state is z be $\mu(z)$, the first-order conditions for c, k', and $a'(z')$ are

$$\mathbf{U}'(c) = \mu(z), \tag{20.A.2}$$

$$\beta \int \mathbf{V}_1(k', a(z'), X') f(z'|z) dz' = \mu(z), \tag{20.A.3}$$

and

$$\beta \mathbf{V}_2(k', a(z'), X') f(z'|z) = \mu(z) Q(z', X). \tag{20.A.4}$$

The envelope conditions are

$$\mathbf{V}_1(k, a(z), X) = \mu(z)(1 + r(X) - \delta) \tag{20.A.5}$$

and

$$\mathbf{V}_2(k, a(z), X) = \mu(z). \tag{20.A.6}$$

Combining (20.A.2), (20.A.4), and (20.A.6), we obtain

$$Q(z', X) = \beta f(z'|z) \frac{\mathbf{U}'(c')}{\mathbf{U}'(c)}, \tag{20.A.7}$$

which is (20.38).

20.A.6 Derivation of $J(X)$, $V(X)$, $W(X)$, and $U(X)$ equations in Section 20.8.1

To derive the values of jobs and workers, we explicitly allow the trade of the claims to profits and wages. Let the quantity of the claim for the profit at the beginning of period t be q_t^J. In the beginning of the period t, the asset (claims) trading occurs with the price $J(X)$, and the new level of asset is $\hat{q}_t^J$. The profit is distributed proportional to $\hat{q}_t^J$. In the next period, some jobs stay matched and some jobs become vacant. Similar notations are used for $V(X)$, $W(X)$, and $U(X)$. Note that q_t^J, $\hat{q}_t^J$, q_t^W, and $\hat{q}_t^W$ each sum up to $(1 - u_t)$, q_t^U and $\hat{q}_t^U$ each sum up to u_t, and q_t^U and $\hat{q}_t^U$ each sum up to $v_t = \theta_t u_t$. Because the consumers (families) are

homogeneous and total number of families is one, the equilibrium values of asset holdings are equal to the corresponding sum.

The budget constraint now becomes (also incorporating the Arrow securities, as in Appendix 20.A.5)

$$c_t + k_{t+1} + \int Q_t(z_{t+1}) a_{t+1}(z_{t+1}) dz_{t+1} + (\hat{q}_t^J - q_t^J) J_t + (\hat{q}_t^V - q_t^V) V_t + (\hat{q}_t^W - q_t^W) W_t + (\hat{q}_t^U - q_t^U) U_t$$
$$= (1 + r_t - \delta) k_t + \hat{q}_t^W w_t + \hat{q}_t^U b + \hat{q}_t^J (y_t - w_t) - \hat{q}_t^V \kappa + a_t(z_t).$$

Considering the definition of d_t (see (20.47)), we can show that the budget constraint in equilibrium is identical to the baseline model and therefore the equilibrium allocation is not altered by the possibility of trading claims.

The transition equations for the asset holdings are

$$q_{t+1}^J = \lambda_f(\theta_t) \hat{q}_t^V + (1 - \sigma) \hat{q}_t^J,$$
$$q_{t+1}^V = (1 - \lambda_f(\theta_t)) \hat{q}_t^V + \sigma \hat{q}_t^J,$$
$$q_{t+1}^W = \lambda_w(\theta_t) \hat{q}_t^U + (1 - \sigma) \hat{q}_t^W,$$

and

$$q_{t+1}^U = (1 - \lambda_w(\theta_t)) \hat{q}_t^U + \sigma \hat{q}_t^W.$$

As in Appendix 20.A.5, we can write down the dynamic programming problem, now with new state variables (which includes the claim holdings) and new constraints above. The value function is now $\mathbf{V}(q^J, q^V, q^W, q^U, k, a(z), X)$. The problem is

$$\mathbf{V}(q^J, q^V, q^W, q^U, k, a(z), X) = \max_{c, k', \{a'(z')\}, \{q^{i\prime}, \hat{q}^i\}^{i=J,V,W,U}} \mathbf{U}(c) + \beta \int \mathbf{V}(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X') f(z'|z) dz'$$

subject to

$$c + k' + \int Q(z', X) a'(z') dz' + (\hat{q}^J - q^J) J(X) + (\hat{q}^V - q^V) V(X) + (\hat{q}^W - q^W) W(X) + (\hat{q}^U - q^U) U(X)$$
$$= (1 + r(X) - \delta) k + \hat{q}^W w(X) + \hat{q}^U b + \hat{q}^J (y(X) - w(X)) - \hat{q}^V \kappa + a(z),$$

$$q^{J\prime} = \lambda_f(\theta(X)) \hat{q}^V + (1 - \sigma) \hat{q}^J,$$
$$q^{V\prime} = (1 - \lambda_f(\theta(X))) \hat{q}^V + \sigma \hat{q}^J,$$
$$q^{W\prime} = \lambda_w(\theta(X))) \hat{q}^U + (1 - \sigma) \hat{q}^W,$$
$$q^{U\prime} = (1 - \lambda_w(\theta(X))) \hat{q}^U + \sigma \hat{q}^W.$$
$$K' = \Omega(X),$$

and

$$u' = (1 - \lambda_w(\theta(X)) + \sigma(1 - u).$$

Let the Lagrange multiplier of the budget constraint be μ and the transition equations be ν^J, ν^V, ν^W, and ν^U. The first-order conditions on c, k', $a'(z')$, and the envelope conditions

on k and $a(z)$ are the same as in Appendix 20.A.5 (i.e., (20.A.2) to (20.A.6)). As a result, we obtain (20.A.7).

The first-order conditions for $\hat{q}^J$, $\hat{q}^V$, $\hat{q}^W$, and $\hat{q}^U$ are:

$$\mu J(X) = \mu(y(X) - w(X)) + \nu^J(1-\sigma) + \nu^V\sigma, \tag{20.A.8}$$

$$\mu V(X) = -\mu\kappa + \nu^J\lambda_f(\theta(X)) + \nu^V(1-\lambda_f(\theta(X))), \tag{20.A.9}$$

$$\mu W(X) = \mu w(X) + \nu^W(1-\sigma) + \nu^U\sigma, \tag{20.A.10}$$

and

$$\mu U(X) = \mu b + \nu^W\lambda_w(\theta(X)) + \nu^U(1-\lambda_w(\theta(X))). \tag{20.A.11}$$

The first-order conditions for $q^{J\prime}$, $q^{V\prime}$, $q^{W\prime}$, and $q^{U\prime}$ are:

$$\nu^J = \beta\int \mathbf{V}_1(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{20.A.12}$$

$$\nu^V = \beta\int \mathbf{V}_2(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{20.A.13}$$

$$\nu^W = \beta\int \mathbf{V}_3(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{20.A.14}$$

$$\nu^U = \beta\int \mathbf{V}_4(q^{J\prime}, q^{V\prime}, q^{W\prime}, q^{U\prime}, k', a(z'), X')f(z'|z)dz'. \tag{20.A.15}$$

Envelope conditions for q^J, q^V, q^W, and q^U are:

$$\mathbf{V}_1(q^J, q^V, q^W, q^U, k, a(z), X) = \mu J(X) \tag{20.A.16}$$

$$\mathbf{V}_2(q^J, q^V, q^W, q^U, k, a(z), X) = \mu V(X) \tag{20.A.17}$$

$$\mathbf{V}_3(q^J, q^V, q^W, q^U, k, a(z), X) = \mu W(X) \tag{20.A.18}$$

$$\mathbf{V}_4(q^J, q^V, q^W, q^U, k, a(z), X) = \mu U(X) \tag{20.A.19}$$

Combining (20.A.8), (20.A.12), (20.A.13), (20.A.16), and (20.A.17) and utilizing (20.A.2) and (20.A.7), we can obtain

$$J(X) = y(X) - w(X) + \int Q(z', X)[(1-\sigma)J(X') + \sigma V(X')]dz',$$

which is (20.39) in the main text. We can similarly obtain (using equations (20.A.8) to (20.A.19))

$$V(X) = -\kappa + \int Q(z', X)[\lambda_f(\theta(X))J(X') + (1-\lambda_f(\theta(X))V(X')]dz',$$

$$W(X) = w(X) + \int Q(z', X)[(1-\sigma)W(X') + \sigma U(X')]dz',$$

and

$$U(X) = b + \int Q(z', X)[\lambda_w(\theta(X))W(X') + (1-\lambda_w(\theta(X)))U(X')]dz'.$$

20.A.7 Calibration and computation of Section 20.8

The parameter values in Table 20.1 apply in this model, except for b. Below, parameters b and κ are endogenously calibrated. First, with $\bar{\theta} = 1$, the steady-state unemployment rate

$$\bar{u} = \frac{\sigma}{\chi + \sigma}.$$

From the Euler equation of the family's consumption-saving problem,

$$\bar{r} = \frac{1}{\beta} - 1 + \delta.$$

Because

$$\bar{r} = \alpha \left(\frac{\bar{K}}{1 - \bar{u}} \right)^{\alpha - 1}$$

can be solved for $\bar{K}$:

$$\bar{K} = \left(\frac{r}{\alpha} \right)^{\frac{1}{1-\alpha}} (1 - \bar{u}).$$

Once we know $\bar{K}$, we can compute

$$\bar{y} = (1 - \alpha) \left(\frac{K}{1 - u} \right)^{\alpha}.$$

The parameter b is set by

$$b = 0.4\bar{y}.$$

From the job creation condition, κ is calibrated as

$$\kappa = \beta\chi(\bar{y} - b) \left(1 - \beta \frac{1 - \sigma - \gamma\chi}{1 - \gamma} \right)^{-1}.$$

Now we can compute the steady-state values of wage and dividend:

$$\bar{w} = \gamma(\bar{y} - b) + b + \gamma\kappa$$

$$\bar{d} = (1 - \bar{u})(y - \bar{w}) - \kappa\bar{u}.$$

22.A Appendix to Chapter 22

22.A.1 Derivation of Equation (22.8)

As a preparation, note two facts about the normal distribution and the lognormal distribution. First, when X is normally distributed $X \sim N(\mu, \sigma^2)$, The random variable αX also follows a normal distribution $N(\alpha\mu, \alpha^2\sigma^2)$. Second, when X is normally distributed $X \sim N(\mu, \sigma^2)$, $\exp(X)$ is lognormally distributed, and $\mathbb{E}[\exp(X)] = \exp(\mu + \sigma^2/2)$.

From (22.6),

$$\ln(A) = (1-\gamma)\ln\left(\int a_i^{\frac{1}{1-\gamma}} di\right).$$

Because $\ln(a_i)$ is normally distributed with $N(\nu - \sigma^2/2, \sigma^2)$, $\ln(a_i)/(1-\gamma)$ is also normally distributed (with mean $(\nu - \sigma^2/2)/(1-\gamma)$ and variance $\sigma^2/(1-\gamma)^2$) and

$$a_i^{\frac{1}{1-\gamma}} = \exp\left(\frac{1}{1-\gamma}\ln(a_i)\right)$$

is lognormally distributed, with mean

$$\exp\left(\frac{1}{1-\gamma}\left(\nu - \frac{\sigma^2}{2}\right) + \frac{1}{2}\frac{\sigma^2}{(1-\gamma)^2}\right) = \exp\left(\frac{1}{1-\gamma}\left(\nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2\right)\right).$$

Because a_i is i.i.d., from the law of large numbers,

$$\int a_i^{\frac{1}{1-\gamma}} di = \mathbb{E}\left[a_i^{\frac{1}{1-\gamma}}\right] = \exp\left(\frac{1}{1-\gamma}\left(\nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2\right)\right),$$

where $\mathbb{E}[\cdot]$ represents expected value. Going back to the first equation, this outcome implies

$$\ln(A) = \nu + \frac{\gamma}{1-\gamma}\frac{1}{2}\sigma^2,$$

and thus (22.8) follows.

22.A.2 Firms versus establishments in the size statistics

Figure 22.7 shows that the fraction of workers working in large firms has increased since the 1990s. The same pattern does not hold for large establishments. Therefore, making a distinction between a firm and an establishment is important in this context.

The BDS dataset does not provide the 10,000+ category for establishments. First, we confirm that the same pattern as Figure 22.7 holds when the threshold moves to 1,000. Figure 22.A.1 plots the fraction of employees working at 1,000+ employee firms. One can see that the graph is (aside from the shift in level) almost identical to Figure 22.7.

Figure 22.A.2 computes the same statistics for 1,000+ employee establishments. Although we see some increase after the mid-2000s, overall the profile has been relatively flat. Thus, the concentration of employment at the top is a firm phenomenon and not an establishment phenomenon.

The contrast between Figures 22.A.1 and 22.A.2 leads us to suspect that it is the number of establishments that is contributing the concentration of employees at very large firms. Figure 22.A.3 confirms this to be the case. It shows that the fraction of establishments that belong to very big firms (10,000+ employees) has steadily increased.

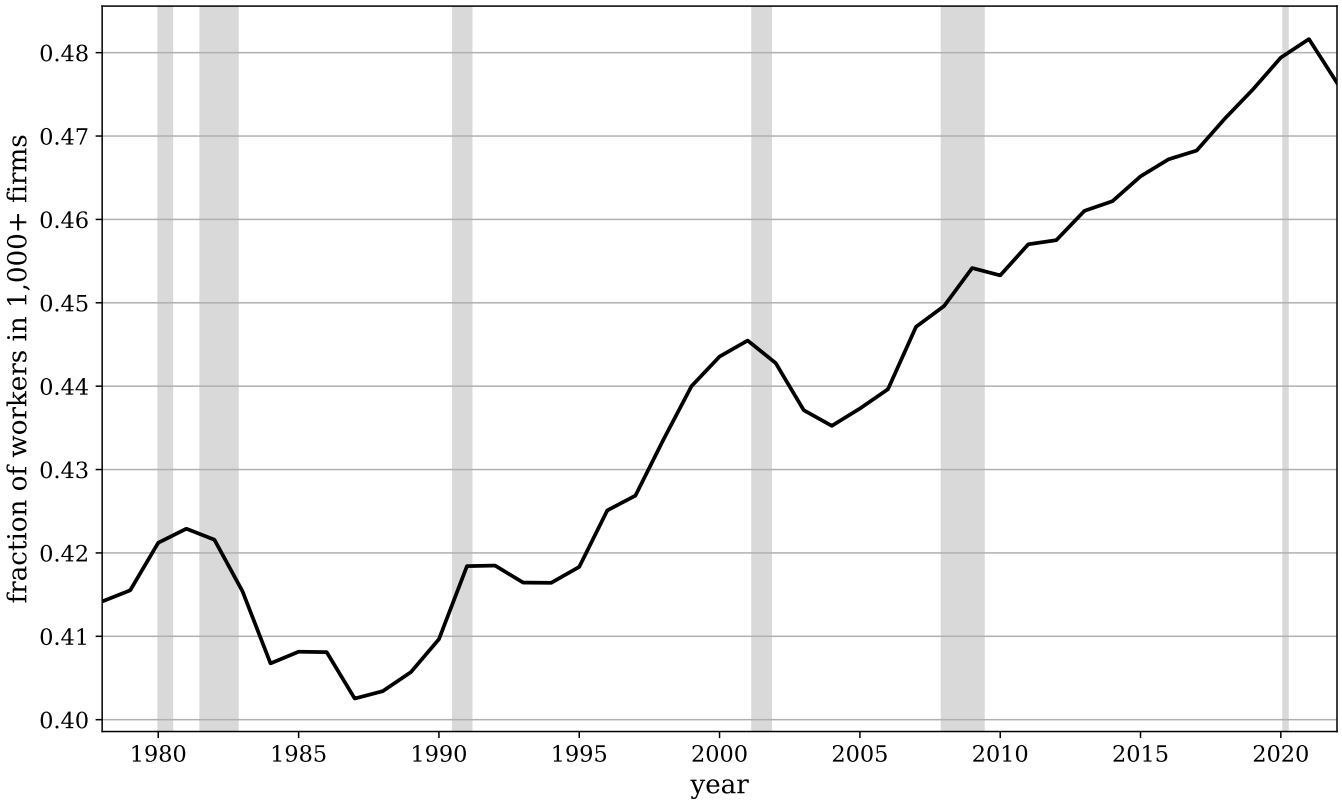


Figure 22.A.1: Fraction of employees working at 1,000+ employee firms.
Source: Business Dynamics Statistics

22.A.3 Derivation of Equation (22.11)

As in Appendix 22.A.1, let us first prepare with a basic property of the normal distribution. When X and Y are jointly normally distributed with $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mu_x, \mu_y)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

Then $\alpha X + \beta Y$ also follows a normal distribution with mean $\alpha\mu_x + \beta\mu_y$ and variance $\alpha^2\sigma_x^2 + \beta^2\sigma_y^2 + 2\alpha\beta\rho\sigma_x\sigma_y$.

From (22.10),

$$\ln(A) = \ln\left(\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}di\right) - \gamma\ln\left(\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{1}{1-\gamma}}di\right).$$

Because $\ln(a_i)$ and $\ln(1-\tau_i)$ follow a bivariate normal distribution,

$$\frac{1}{1-\gamma}\ln(a_i) + \frac{\gamma}{1-\gamma}\ln(1-\tau_i)$$

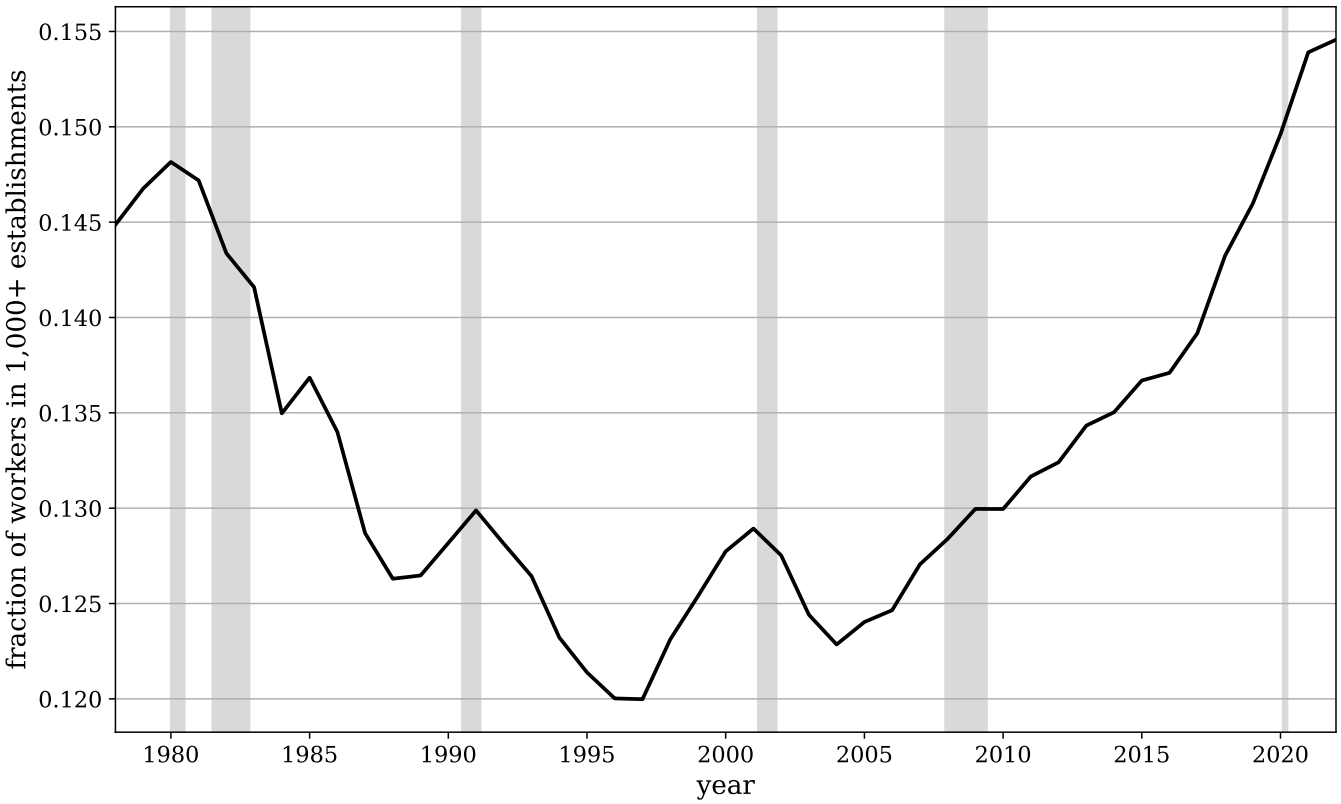


Figure 22.A.2: Fraction of employees working at 1,000+ employee establishments. Source: Business Dynamics Statistics

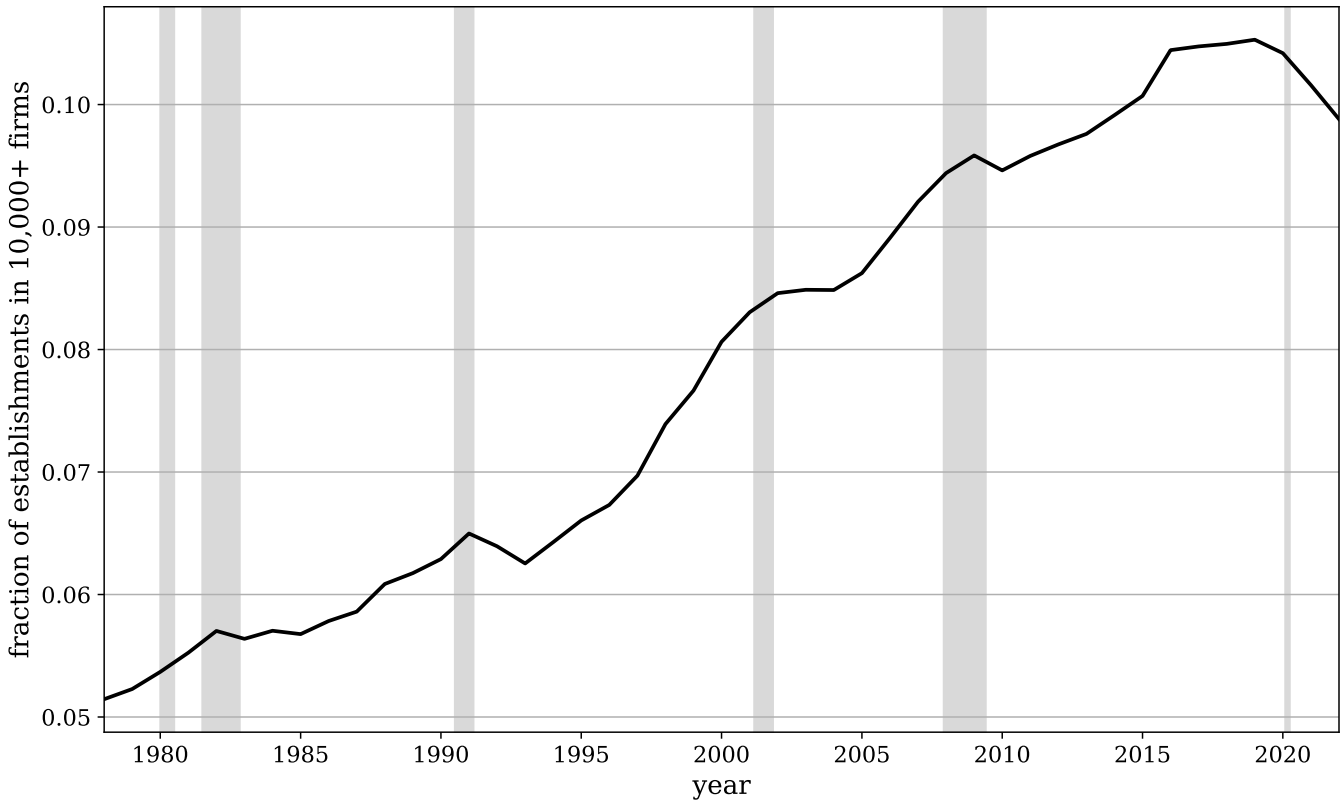


Figure 22.A.3: Fraction of establishments that belong to 10,000+ employee firms. Source: Business Dynamics Statistics

follow a normal distribution with mean

$$\frac{1}{1-\gamma}\left(\nu_a - \frac{\sigma_a^2}{2}\right) + \frac{\gamma}{1-\gamma}\left(\nu_\tau - \frac{\sigma_\tau^2}{2}\right)$$

and variance

$$\frac{1}{(1-\gamma)^2}\sigma_a^2 + \frac{\gamma^2}{(1-\gamma)^2}\sigma_\tau^2 + 2\rho\frac{\gamma}{(1-\gamma)^2}\sigma_a\sigma_\tau.$$

From the law of large numbers,

$$\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}di = \mathbb{E}\left[a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{\gamma}{1-\gamma}}\right] = \mathbb{E}\left[\exp\left(\frac{1}{1-\gamma}\ln(a_i) + \frac{\gamma}{1-\gamma}\ln(1-\tau_i)\right)\right]$$

holds, and it can be computed as

$$\exp\left(\frac{1}{1-\gamma}\left(\nu_a - \frac{\sigma_a^2}{2}\right) + \frac{\gamma}{1-\gamma}\left(\nu_\tau - \frac{\sigma_\tau^2}{2}\right) + \frac{1}{2}\left(\frac{1}{(1-\gamma)^2}\sigma_a^2 + \frac{\gamma^2}{(1-\gamma)^2}\sigma_\tau^2 + 2\rho\frac{\gamma}{(1-\gamma)^2}\sigma_a\sigma_\tau\right)\right).$$

We can similarly compute

$$\int a_i^{\frac{1}{1-\gamma}}(1-\tau_i)^{\frac{1}{1-\gamma}}di$$

and after some algebra, we obtain

$$\ln(A) = \nu_a + \frac{\gamma}{1-\gamma}\frac{1}{2}(\sigma_a^2 - \sigma_\tau^2),$$

implying (22.11).

22.A.4 Derivation of the Bertrand competition result in Section 22.6.2

First, as in the monopolistic competition case, consider the cost-minimization problem for the final-good producer. Within a sector, it has to solve

$$\min_{\{q_{ij}\}} \sum_{j=1}^{J} \hat{p}_{ij}q_{ij}$$

subject to

$$y_i = \left[\sum_{j=1}^{J} q_{ij}^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}}$$

for given y_i. The first-order condition is

$$\hat{p}_{ij} = \lambda_i q_{ij}^{-\frac{1}{\eta}} y_{ij}^{\frac{1}{\eta}}.$$

Similarly to the monopolistic competition case, we can think of λ_i as the price of the combined good for sector i (call it p_i) and

$$p_i = \left[\sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta}\right]^{\frac{1}{1-\eta}}.$$

An intermediate-good producer with Bertrand competition therefore solves

$$\max_{\hat{p}_{ij}} \ \hat{p}_{ij} q_{ij} - cm_{ij},$$

given

$$q_{ij} = a_{ij} m_{ij}^{\gamma} \tag{22.A.1}$$

and

$$\hat{p}_{ij} = p_i q_{ij}^{-\frac{1}{\eta}} y_{ij}^{\frac{1}{\eta}}. \tag{22.A.2}$$

Now, in addition to these, the producer is aware that its price $\hat{p}_{ij}$ affects the sectoral price p_i and thus the sectoral demand. Thus

$$p_i = y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} \tag{22.A.3}$$

(which is the solution of the cost-minimization problem across sectors, with the price index normalized to one) and

$$p_i = \left[\sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta}\right]^{\frac{1}{1-\eta}} \tag{22.A.4}$$

are also given to the producer (in addition to Y).

The first-order condition is

$$\hat{p}_{ij} \frac{\partial q_{ij}}{\partial \hat{p}_{ij}} + q_{ij} = \frac{\partial (cm_{ij})}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial \hat{p}_{ij}}$$

and thus

$$\left(1 - \frac{q_{ij}/\hat{p}_{ij}}{\partial q_{ij}/\partial \hat{p}_{ij}}\right) \hat{p}_{ij} = \mathcal{M}.$$

Here, the marginal cost is

$$\mathcal{M} = \frac{\partial (cm_{ij})}{\partial q_{ij}} = \frac{cm_{ij}^{1-\gamma}}{\gamma a_{ij}},$$

as in the case with the Cournot competition, computed using (22.A.1). Using (22.A.2), (22.A.3), and (22.A.4),

$$\frac{q_{ij}/\hat{p}_{ij}}{\partial q_{ij}/\partial \hat{p}_{ij}} = \frac{1}{\varepsilon(s_{ij})},$$

where

$$\varepsilon(s_{ij}) = \eta(1 - s_{ij}) + \sigma s_{ij}.$$

Thus

$$\hat{p}_{ij} = \frac{\varepsilon(s_{ij})}{\varepsilon(s_{ij}) - 1} \mathcal{M}.$$

22.A.5 Proof of Hulten's theorem

This Appendix outlines the proof of Hulten's (1978) theorem, partly following Baqaee and Farhi (2019). The theorem to prove is

$$\frac{dY}{Y} = \sum_i D_i \frac{da_i}{a_i},$$

where D_i is the Domar weight:

$$D_i = \frac{p_i y_i}{\sum_i p_i c_i}.$$

Below we show the theorem in a relatively simple static economy. The only input is labor, and labor is inelastically supplied. Suppose that there are N goods. Assume that all markets are competitive. The representative consumer's utility takes the form

$$U(c_1, ..., c_N)$$

and the budget constraint is

$$\sum_{i=1}^{N} p_i c_i = w\bar{\ell} + \sum_{i=1}^{N} \pi_i,$$

where w is the wage rate and $\bar{\ell}$ is the fixed amount of labor supply, which is the only input for production. c_i is the consumption of good i, and π_i is profit from sector i. Assume that the preferences are homothetic so that $U(c_1, ..., c_N)$ is linearly homogeneous.

The production function for sector i is

$$y_i = a_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN}),$$

where ℓ_i is labor input at sector i and x_{ij} is the quantity of product j used in sector i. The profit is

$$\pi_i = p_i y_i - w\ell_i - \sum_{j=1}^{N} p_j x_{ij}.$$

The market-clearing conditions are

$$y_i = \sum_{j=1}^{N} x_{ji} + c_i$$

for all i and

$$\bar{\ell} = \sum_{i=1}^{N} \ell_i.$$

First, consider the consumer's expenditure minimization problem for a given utility level

$$\min_{c_1, ... c_N} \sum_{i=1}^{N} p_i c_i$$

subject to

$$U(c_1, ..., c_N) = u.$$

The Lagrangian is

$$L = \sum_{i=1}^{N} p_i c_i - \lambda(U(c_1, ..., c_N) - u).$$

The first-order condition for this problem is

$$p_i = \lambda \frac{\partial U(c_1, ..., c_N)}{\partial c_i}. \tag{22.A.5}$$

Let us normalize the prices (i.e., choose the numeraire) p_i so that $\lambda = 1$ in equilibrium. Let

$$Y \equiv \sum_{i=1}^{N} p_i c_i$$

be the GDP (and TFP) of this economy. From (22.A.5) and linear homogeneity, it can be rewritten as

$$Y = \sum_{i=1}^{N} \frac{\partial U(c_1, ..., c_N)}{\partial c_i} c_i = U(c_1, ..., c_N). \tag{22.A.6}$$

Therefore, in equilibrium, the level of utility also represents GDP.

Because the first welfare theorem holds, the competitive equilibrium is Pareto optimal, and solves the social planner's problem

$$\max_{c_i, x_{ij}, \ell_i} U(c_1, ..., c_N)$$

subject to

$$c_i + \sum_{j=1}^{N} x_{ji} = a_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN})$$

and

$$\sum_{i=1}^{N} \ell_i = \bar{\ell}.$$

The Lagrangian for the social planner is

$$L = U(c_1, ..., c_N) + \sum_{i=1}^{N} \mu_i \left(a_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN}) - c_i - \sum_{j=1}^{N} x_{ji} \right) + \nu \left(\bar{\ell} - \sum_{i=1}^{N} \ell_i \right).$$

From the first-order condition,

$$\frac{\partial U(c_1, ..., c_N)}{\partial c_i} = \mu_i. \tag{22.A.7}$$

The envelope theorem implies

$$\frac{dU}{da_i} = \mu_i F_i(\ell_i, x_{i1}, x_{i2}, ..., x_{iN}) = \mu_i y_i \frac{1}{a_i}.$$

From (22.A.6), this equation implies

$$\frac{dY}{da_i} = \mu_i y_i \frac{1}{a_i}. \tag{22.A.8}$$

From (22.A.5) and (22.A.7), together with our normalization of $\lambda = 1$,

$$\mu_i = p_i$$

holds. Using this relationship and dividing both sides of (22.A.8) by $Y = \sum_{i=1}^{N} p_i c_i$ yields

$$\frac{dY}{Y} = \frac{p_i y_i}{\sum_{i=1}^{N} p_i c_i} \frac{da_i}{a_i}.$$

Repeating the same procedure for all a_i, we obtain the theorem.

22.A.6 Firm size distribution in Section 22.8

Let the mass of firms with k product lines at time t be M_{kt}. The transition equations for M_{kt} are as follows.

First, for M_{1t}, there are three kinds of firms in $M_{1,t+1}$ next period. First is the entrants, second is the one-product firms that remain with one product, and third is the multi-product firms who had more than one product but lost some product lines and became one-product firms.

$$M_{1,t+1} = \nu_t + M_{1,t} \sum_{i=0}^{1} \mathbf{P}_\mu(1,i)\mathbf{P}_\eta(1,i) + \sum_{h=1}^{\infty} \left(M_{1+h,t} \sum_{i=h}^{1+h} \mathbf{P}_\mu(1+h,i)\mathbf{P}_\eta(1+h,i-h) \right),$$

where

$$\mathbf{P}_\mu(a,b) = \left(\begin{array}{c} a \\ b \end{array} \right) \mu^b (1-\mu)^{a-b}$$

is the probability of losing i product lines when starting from k lines, and

$$\mathbf{P}_\eta(a,b) = \left(\begin{array}{c} a \\ b \end{array} \right) \eta^b (1-\eta)^{a-b}$$

is the probability of gaining i product lines when starting from k lines.

Second, for M_{kt} for $k > 1$, there are three different types in $M_{k,t+1}$. First is the firms starting from $k-h$ products, with a net gain of h products and getting to k. The second is the k-product firms that remain with k products. The third is firms with $k+h$ products, where $h > 0$ and that have a net loss of h products and get to k products.

$$\begin{aligned} M_{k,t+1} = & \sum_{h=1}^{\lfloor k/2 \rfloor} \left(M_{k-h,t} \sum_{i=0}^{\lfloor k/2-h \rfloor} \mathbf{P}_\mu(k-h,i)\mathbf{P}_\eta(k-h,i+h) \right) \\ & M_{k,t} \sum_{i=0}^{k} \mathbf{P}_\mu(k,i)\mathbf{P}_\eta(k,i) + \sum_{h=1}^{\infty} \left(M_{k+h,t} \sum_{i=h}^{k+h} \mathbf{P}_\mu(k+h,i)\mathbf{P}_\eta(k+h,i-h) \right). \end{aligned}$$

Here, $\lfloor x \rfloor$ represents the integer not exceeding x. To find the stationary distribution of firm size, we can look for $\{M_1, M_2, ...\}$ such that $M_{1,t+1} = M_{1,t}$, $M_{2,t+1} = M_{2,t}$, ... hold.

24.A Appendix to Chapter 24

We begin with the numerical computation and then discuss calibration.

24.A.1 Computation

The endogenous bond price q introduces computational challenges absent in the neoclassical growth model. We first describe the simplest computation strategy to solve the model but two computational challenges that arise with these strategy. We then present alternative computational strategies.

The simplest computational strategy is to define a grid for debt levels $\mathcal{B} = \left\{B^1, ..., B^M\right\}$, income levels $\mathcal{Y} = \left\{Y^1, ..., Y^N\right\}$, and transition probabilities $\pi_{i,j} = Pr\left(Y_{t+1} = Y^j \mid Y_t = Y^i\right)$. Then, an equilibrium is found by:

1. Starting with an initial guess for $\left\{V_0^R, V_0^D, V_0, q_0, \hat{B}_0, \hat{D}_0\right\}$.

2. Updating V^R and $\hat{B}$ in iteration I by solving

$$V_{I+1}^R(B^h, Y^i) = \max_{B' \in \mathcal{B}} \left\{ u\left(C\right) + \beta \sum_{j=1}^{N} V_I(B', Y^j)\pi_{i,j} \right\} \tag{24.A.1}$$
$$s.t. \quad C = Y^i - \delta B^h + q_I(B', Y^i)\left(B' - (1-\delta)B^h\right) \quad \forall h = 1, ..., M, i = 1, ..., N.$$

Updating V^D in iteration I by computing

$$V_{I+1}^D(Y^i) = u\left(Y^i - \phi\left(Y^i\right)\right) + \beta \sum_{j=1}^{N} \pi_{i,j}\left[\psi V_I(0, Y') + (1-\psi)\, V_I^D(Y')\right] \quad \forall i = 1, ..., N.$$

Updating V and $\hat{D}$ in iteration I by computing

$$V_{I+1}\left(B^h, Y^i\right) = \max_{d \in \{0,1\}} \{D V_{I+1}^D(Y^i) + (1-D) V_{I+1}^R(B^h, Y^i)\} \quad \forall h = 1, ..., M, i = 1, ..., N.$$

Updating q in iteration I by computing

$$q_{I+1}(B', Y^i) = \frac{1}{1+r} \sum_{j=1}^{N} \left[1 - \hat{D}_I\left(B', Y^j\right)\right] \left[\delta + (1-\delta) q_I(\hat{B}_I(B', Y^j), Y^j)\right] \pi_{i,j} \quad \forall B' \in \mathcal{B}, i = 1, ..., N. \tag{24.A.2}$$

3. Verifying convergence for a convergence criterion ϵ. That is, verifying that for all $B^h \in \mathcal{B}$ and $Y^i \in \mathcal{Y}$,

   - $Max\left\{\mid V_{I+1}(B^h, Y^i) - V_I(B^h, Y^i) \mid\right\} < \epsilon$ and

- $Max\left\{\mid q_{I+1}(B^h, Y^i) - q_I(B^h, Y^i) \mid\right\} < \epsilon$.

If both conditions are satisfied, we approximate the equilibrium with

$$\left\{V^R_{I+1}, V^D_{I+1}, V_{I+1}, q_{I+1}, \hat{B}_{I+1}, \hat{D}_{I+1}\right\}.$$

If one of the above conditions is not satisfied, we return to step 2.

Problem (24.A.1) is of straightforward implementation: We just need to find the debt B' that yields the highest value in a vector of possible continuation values. However, it can generate significant approximation errors if the number of grid points M and N are not sufficiently large (Hatchondo, Martinez, and Sapriza, 2010).

Assuming that the true functions are differentiable, the total derivative of the equilibrium price with respect to aggregate income is represented by

$$\frac{dq(\hat{B}(B,Y),Y)}{dY} = q_1(\hat{B}(B,Y),Y)\frac{\hat{B}(B,Y)}{\partial Y} + q_2(\hat{B}(B,Y),Y), \tag{24.A.3}$$

where $q_1(q_2)$ denotes the derivative of the bond price schedule with respect to the first (second) argument. If the number of grid points for debt choices is not large enough, the solution of problem (24.A.1) could generate the same debt choice B' for two consecutive income realizations (Y^n, Y^{n+1}), thus "shutting down" the effect represented by the first term on the right-hand side of equation (24.A.3). This approximation error is important because it distorts the volatility of the spread and the co-movement between the spread and income implied by the model (Hatchondo et al., 2010).

The second computational challenge involves the approximation of the bond price schedule itself. First, if there is a point on the grid (B^m, Y^n) such that the sovereign is nearly indifferent between repaying and defaulting, numerical approximation could generate that the default decision $\hat{D}(B^m, Y^n)$ jumps from 0 to 1 (or vice versa) from one iteration to next. Those jumps in the default decision would have a sizable effect on the bond price schedule for income realizations i with high transition probabilities $\pi_{i,n}$. This could prevent the bond price from converging. This problem is more severe when the number of grid points for income realizations is not large enough.

Second, the objective function in problem (24.11) may not be concave and may have multiple local maxima. If there is a state (B^h, Y^i) with two local maxima $\underline{B}$ and $\bar{B}$ that yield a similar value of repaying V^R, approximation error could induce the code to cycle between those two local maxima from one iteration to the next. Since both local maxima generate nearly the same continuation value under repaying, these jumps may not compromise the convergence of V^R. However, these jumps would typically compromise convergence of the bond price q as the higher local maximum implies a higher future default probability.

In sum, a large number of grid points for income realizations is necessary for avoiding convergence problems, and for the model to capture more accurately the direct and indirect effects of income on the bond prices in equation (24.A.2). Below, we summarize computational alternatives to deal with the above challenges.

Method 1. Instead of solving for $B' \in \mathcal{B}$ in equation (24.A.1), we could use numerical optimization routines to solve for $B' \geq 0$. In addition, instead of approximating the probability distribution for Y_t with a discrete process, we could use quadrature methods to approximate the expected values $\mathbb{E}\left[V_I(B', Y') \mid Y\right]$ and $\mathbb{E}\left[(1 - \hat{D}_I(B', Y'))\left[\delta + (1-\delta)q_I(\hat{B}_i(B', Y'), Y')\right] \mid Y\right]$. The first expectation is necessary to compute the value of repaying V^R_{I+1}, while the second expectation is necessary to evaluate the bond price $q_I(B', Y^i)$ on debt choices that are not on the grid.[37] Both computational strategies require using interpolation methods to evaluate $V_I(B', Y')$ and $q_I(B', Y')$ on states (B', Y') that lie outside the grid $\mathcal{B} \times \mathcal{Y}$.

The results presented in this chapter were computed by approximating the value functions V^R and V^D, and the bond price function q using linear interpolation over income (Y) and cubic spline interpolation over debt (B). We use 30 grid points for debt and income. Expectations are computed using 100 quadrature points for the income shock. A large number of quadrature points minimizes the effects of potential cycles in the default or borrowing decisions. We solve for the optimal borrowing problem in each state by i) searching over a grid of debt levels B' and ii) using the optimal B' on that grid as an initial guess in a nonlinear optimization routine. The initial guess corresponds to the solution in the final period of a finite horizon model:

$$
\begin{aligned}
&V^R_0(B, Y) = u(Y - \delta B), \\
&V^D_0(Y) = u(Y - \phi(Y)), \\
&q_0(B', Y) = 0, \\
&\hat{B}_0(B, Y) = 0, \text{ and} \\
&\hat{D}_0(B, Y) = \begin{cases} 1 & \text{if } V^D_0(Y) \text{ and } > V^R_0(B, Y) \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

In the last period of the finite horizon game, the sovereign has no access to credit markets and consumes all available resources.[38] We iterate using a convergence criterion of $\varepsilon = 10^{-6}$.

Method 2. Chatterjee and Eyigungor (2012) assume that the economy lives on the space $\mathcal{B} \times \mathcal{Y}$ and that aggregate income is also affected by a purely temporary income shock m with a continuous distribution. Formally, the value of defaulting has the following recursive formulation

$$
V^D(Y^i, m) = u\left(Y^i - \phi\left(Y^i\right) - m\right) + \beta \mathbb{E}_{m'} \sum_{j=1}^{N} \left[\psi V(0, Y^j, m') + (1 - \psi) V^D(Y^j, m')\right] \pi_{i,j},
$$

[37] We can approximate the default rule $\hat{D}_I(B', Y') = \mathbb{1}_{V^D_I(Y') > V^R_I(B', Y')}$, where $\mathbb{1}$ denotes the indicator function. Quadrature methods approximate

$$
\int_a^b f(x)dx \simeq \sum_{i=1}^{L} \omega_i f(x_i).
$$

[38] Hatchondo et al. (2010) discuss the advantages of using interpolation and solving for the equilibrium of a finite-horizon economy.

where $\mathbb{E}_{m'}$ denotes the expectation over temporary next-period shocks. The value of repaying has the following formulation

$$V^R(B^h, Y^i, m) = \max_{B' \in \mathcal{B}} \left\{ u(C) + \beta \mathbb{E}_{m'} \sum_{j=1}^{N} V(B', Y^j, m') \pi_{i,j} \right\} \tag{24.A.4}$$
$$s.t. \quad C = Y^i - \delta B^h + q(B', Y^i) \left[B' - (1-\delta) B^h \right].$$

The advantage of this approach is that the continuous income shock smoothes out the problem: Any jump in the default or borrowing decisions in state (B^m, Y^n, m) has zero probability mass and, consequently, does not affect the bond price.[39]

Method 3. Dvorkin, Sanchez, Sapriza, and Yurdagul (2021) present an alternative strategy to smooth out the problem by resorting to extreme value shocks. They assume that the continuation values under repayment and default are subject to shocks. Formally, they assume that each period, the economy is hit with $M+1$ shocks encapsulated in the vector $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_M, \epsilon_{M+1})$. The ϵ shocks follow an i.i.d. process and are drawn from a multivariate distribution with continuous support. The shock ϵ_{M+1} affects the continuation value after defaulting:

$$V^D(Y^i, \epsilon_{M+1}) = u\left(Y^i - \phi\left(Y^i\right)\right) + \beta \sum_{j=1}^{N} \left[\psi V(0, Y^j, \boldsymbol{\epsilon}') + (1-\psi) V^D(Y^j, \epsilon'_{M+1}) \right] \pi_{i,j} + \epsilon_{M+1}.$$

The shocks $(\epsilon_1, ..., \epsilon_M)$ affect the continuation value of repayment conditional on the bond position chosen by the sovereign, i.e.,

$$V^R(B^h, Y^i, \boldsymbol{\epsilon}) = \max_{h'=1,...,N} \left\{ u(C) + \beta \sum_{j=1}^{N} V\left(B', Y^j, \boldsymbol{\epsilon}'\right) \pi_{i,j} \right\} + \epsilon_{h'} \tag{24.A.5}$$
$$s.t. \quad C = Y^i - \delta B^h + q(B^{h'}, Y^i) \left[B^{h'} - (1-\delta) B^h \right].$$

The sovereign decides whether to default and how much to borrow after observing $\boldsymbol{\epsilon}$.[40] However, ex-ante default and borrowing decisions are stochastic. In effect, the probability that a sovereign defaults after entering a period with initial debt B^h and income Y^i is $D(B^h, Y^i) = \mathbb{E}_{\boldsymbol{\epsilon}} \hat{D}(B^h, Y^i, \boldsymbol{\epsilon})$. The probability that the sovereign chooses debt $B' = B^m$ in a period with debt and income (B^h, Y^i) is denoted by $B(B^m \mid B^h, Y^i)$. That probability depends on ϵ_m taking a sufficiently larger value relative to the other shocks.

Dvorkin et al. (2021) show that the equilibrium bond price satisfies

$$q(B^m, Y^i) = \frac{1}{1+r} \sum_{j=1}^{N} \left[1 - D\left(B^m, Y^j\right)\right] \left[\delta + \sum_{h=1}^{M} (1-\delta) q(B^h, Y^j) B(B^h \mid B^m, Y^j) \right] \pi_{i,j}. \tag{24.A.6}$$

[39] Gordon and Qiu (2018) propose an algorithm that enables to solve problem (24.A.4) without evaluating B' on all the points of the space $\mathcal{B}$ and can significantly speed up convergence.

[40] Our exposition simplifies the setup in Dvorkin et al. (2021). They allow the sovereign to choose a richer portfolio problem and to negotiate with investors after a default (resulting in a positive recovery value of debt in default).

Notably, relative to equation (24.A.2), the above formulation replaces the default decision $\hat{D}(B^m, Y^j)$ with the default probability D, and the portfolio decision $\hat{B}(B^m, Y^j)$ with a probability distribution of possible debt choices. Both features minimize the effects of potential cycles in the default or borrowing rules from one iteration to the next.

Uniqueness. Auclert and Rognlie (2016) and Aguiar and Amador (2019) show that in versions of the model with one-period bonds (i.e., when $\delta = 1$), a unique equilibrium exists. Furthermore, Aguiar and Amador (2019) show that the model can be recast in a formulation that satisfies Blackwell's sufficient conditions. However, Aguiar and Amador (2020) show that there can be multiple Markov equilibria in the model with long-term debt ($\delta \in (0, 1)$), implying that the initial guess plays a non-trivial role in the computation of the equilibrium. By using as initial guesses the equilibrium functions in the last period (T) of a finite-horizon economy, we are finding the equilibrium that is the limit of the finite horizon game as $T \to \infty$.

24.A.2 Calibration

Functional forms. We need to specify functional forms to solve the model numerically. We assume that the utility function displays a constant coefficient of relative risk aversion, i.e.,

$$u\left(C\right) = \frac{c^{1-\sigma} - 1}{1 - \sigma}, \text{ with } \sigma \neq 1.$$

Finally, as in Chatterjee and Eyigungor (2012), we assume a nonlinear income cost of defaulting ϕ:

$$\phi(Y) = max\left\{Y\left[\lambda_0 Y + \lambda_1\left[Y - \mathbb{E}(Y)\right]\right], 0\right\}, \tag{24.A.7}$$

where $\mathbb{E}(Y)$ denotes the unconditional mean income. Equation (24.A.7) assumes a non-negative income cost of defaulting. The parameter λ_0 determines the average fraction of income lost during defaults. The parameter λ_1 determines how sensitive the fraction of income lost during defaults is to the income level. Henceforth, we refer to λ_0 and λ_1 as the average and the slope parameters of the income cost of defaulting, respectively.[41]

Parameters chosen outside the model. We use data from Mexico for choosing the parameters that govern the endowment process, the level and duration of debt, and the mean spread. Mexico is a standard reference in studies of emerging economies and displays the same properties that are observed in other emerging economies (**?**; Neumeyer and Perri, 2005; Uribe and Yue, 2006). A period in the model refers to one quarter. When available, we use quarterly data from 1993 to 2018. Table 24.A.1 presents the benchmark values given to all parameters in the model.

We estimate the parameter values that govern the income process (ρ and σ_ϵ) using the logarithm of detrended GDP per capita in Mexico. The value for μ is chosen so that the unconditional income expectation $\mathbb{E}(Y) = 1$.

[41] Note that when defaults tend to be declared at low-income realizations, the fraction of income lost during defaults observed along the equilibrium path is below λ_0.

Table 24.A.1: Benchmark parameter values.

Income autocorrelation coefficient	ρ	0.91	Mexican data
Standard deviation of innovations	σ_ϵ	1.3%	Mexican data
Mean log income	μ	$(-1/2)\sigma_\epsilon^2$	Assumption
Variable	Parameter	Value	Source
Debt duration	δ	0.033	Mexican data
Risk aversion	σ	2	Literature
Risk-free rate	r	1%	Literature
Discount factor	β	0.975	Literature
Probability exclusion ends	ψ	0.083	Cross-country data
Income cost of defaulting	λ_0	0.184	Debt and spread targets
Income cost of defaulting	λ_1	2.059	Debt and spread targets

The bonds assumed in this chapter do not have a maturity date. Therefore, we choose to parameterize δ to target average duration instead of a maturity measure. We set $\delta = 3.3\%$ to generate an average debt duration 5 years in the simulations.[42] We use the Macaulay definition of duration: the duration of an asset that pays x_t in period t consists of

$$\text{Duration} = \frac{\frac{\sum_{t=0}^{\infty} t x_t}{(1+i)^t}}{\frac{\sum_{t=0}^{\infty} x_t}{(1+i)^t}},$$

where i is the rate used to discount future payoffs. For the bonds assumed in this paper, the duration in period t equals $\frac{1+i_t}{\delta+i_t}$, where i_t denotes the bond yield. Given the target for the average bond yield (risk-free rate plus spread), the duration target pins down the value for δ.

The coefficient of relative risk aversion is set equal to 2, a standard value in real business cycle studies. In the standard neoclassical model without uncertainty, the real interest rate is endogenous and determined by the discount factor β. In this model, the small-open economy assumption manifests in the economy taking as given international prices, such as the global risk-free interest rate. The quarterly risk-free interest rate (r) is set to 1 percent, which is standard in real business cycle studies. Following other quantitative studies, the discount factor β is set to 0.975. A key property of the calibration is that $\beta(1+r) < 1$. Our objective of targeting a negative average asset position can only be attained with a sufficiently impatient sovereign.[43]

Finally, we assume an average duration of post-default exclusion periods of three years, based on the estimations in Dias and Richmond (2009). This implies that the probability of regaining access to debt markets is $\psi = 0.083$.

[42]The data for duration corresponds to the average Modified Duration for Mexican government bonds computed by J.P. Morgan between January 2002 and March 2018.

[43]Using a model without defaults, Aiyagari (1994) shows that due to self-insurance incentives, the demand for assets would be unlimited if $\beta(1+r) = 1$.

Parameters chosen using the model. The parameters λ_0 and λ_1 are jointly calibrated, targeting an average debt-to-GDP ratio and yield spread. For spread data, we use the J.P. Morgan EMBI+ spread from 1994 to 2018. In the model simulations, the bond yield i_t is the return an investor would earn if it holds the bond to maturity (forever) and no default is declared. This yield satisfies

$$q_t = \sum_{j=1}^{\infty} \frac{\delta(1-\delta)^{j-1}}{(1+i_t)^j}.$$

The annualized sovereign spread, r_t^s, is then computed as

$$r_t^s = \left(\frac{1+i_t}{1+r}\right)^4 - 1.$$

How to choose the target for debt is less clear-cut and there is more variation in the debt level targeted in the literature. Some studies target external public debt. Other studies target the unsecured fraction of external public debt, i.e., the fraction of debt that sovereigns do not pay after a default. We target Mexico's average gross public debt ratio from 1996 to 2019, which is 43 percent according to the IMF WEO dataset. To a great extent, what is the appropriate choice of the debt target depends on the application at hand (and may be inconsequential for some applications, including the issues underscored in this chapter). For example, focusing on the total public debt ratios may offer a more palatable laboratory when one wants to emphasize the sovereign's rollover needs. Debt levels in the simulations are calculated as the present value of future payment obligations discounted at the risk-free rate, that is, $\frac{\delta}{\delta+r}B_t$, and we report debt levels as a percentage of annualized income.

Table 24.A.1 shows that we match the calibration targets. The next section explains how we do so.

24.A.3 Matching targeted moments

This section explains how choosing the values for the cost of default parameters allows us to match the calibration targets. The left panel of Figure 24.A.1 shows that the average debt level in the simulations increases with the value of the average cost parameter λ_0 and decreases with the slope parameter λ_1. The right panel of Figure 24.A.1 shows that the average spread in the simulations displays the opposite behavior: It decreases with λ_0 and it increases with λ_1. Therefore, as illustrated in Figure 24.A.2, curves with positive slope can be used to represent both (i) the combination of values of λ_0 and λ_1 that allows the model to generate the targeted debt level, and (ii) the combination of values of λ_0 and λ_1 that allows the model to generate the targeted spread level.[44] Figure 24.A.2 shows a unique combination of parameter values that allows the model to match both targets. Next, we

[44]Let X denote the function that determines the value of moment x in the simulations as a function of the parameters that determine the income cost of defaulting (λ_0, λ_1). The implicit function theorem implies that the combination of parameters (λ_0, λ_1) for which the model simulations replicates the target value for x, denoted by $\bar{X}$, satisfies:

$$\left.\frac{\partial \lambda_1}{\partial \lambda_0}\right|_{X(\lambda_0,\lambda_1)=\bar{X}} = -\frac{\partial X(\lambda_0,\lambda_1)/\partial \lambda_0}{\partial X(\lambda_0,\lambda_1)/\partial \lambda_1}.$$

explain that this is a natural outcome in models of equilibrium default because the average cost parameter λ_0 affects more the debt level and the slope cost parameter λ_1 affects more the spread level.

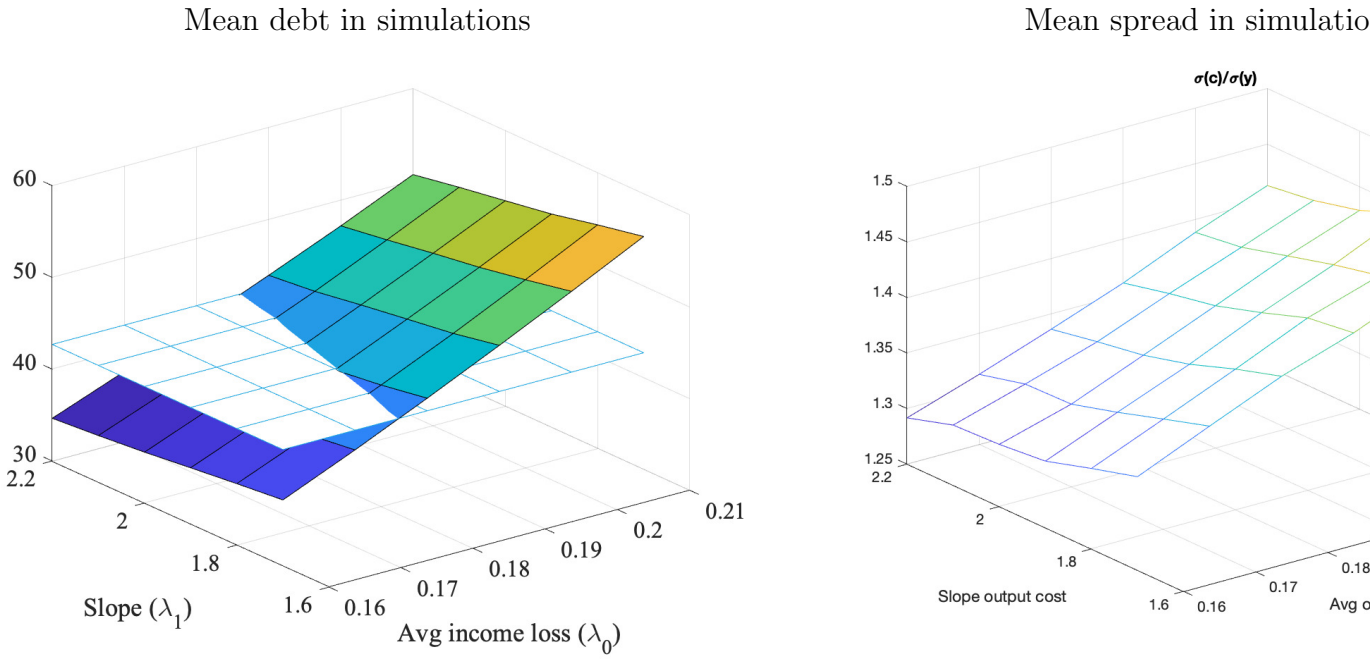


Figure 24.A.1: Average ratio of debt to annual income and annual spread as functions of the income cost parameters. The horizontal planes correspond to the target values for debt and spread.

Role of the parameter λ_0. Figure 24.A.3 illustrates how the average cost parameter λ_0 affects the equilibrium levels of debt and spread (reflected in the equilibrium price at which the sovereign sells bonds). The left panel presents the income threshold at which the sovereign is indifferent between defaulting and repaying.[45] As is standard in equilibrium default models, this threshold is unique. Since the cost of defaulting is an increasing function of income, the sovereign defaults for income realizations below the threshold. The left panel of Figure 24.A.3 shows that an increase in the value of the average cost parameter λ_0 expands the repayment region—i.e., the combination of debt and income levels for which the sovereign would choose to repay its debt. Naturally, there are states where the sovereign chooses to default when the cost is lower and chooses not to default when the cost is higher.

The right panel of Figure 24.A.3 presents the bond price function q in a period in which income is equal to its unconditional mean. That is, the panel presents the menu of debt and bond prices available to the sovereign when it chooses the next period debt level, and in turn, the spread it pays. A higher value of the average cost parameter λ_0 relaxes the sovereign's borrowing constraint. Since it is more costly for the sovereign to default on all debt levels, lenders are willing to pay a higher price for sovereign bonds.

[45] Formally, the left panel of Figure 24.A.3 plots the function Y^* that satisfies $V^R(B, Y^*(B)) = V^D(Y^*(B))$.

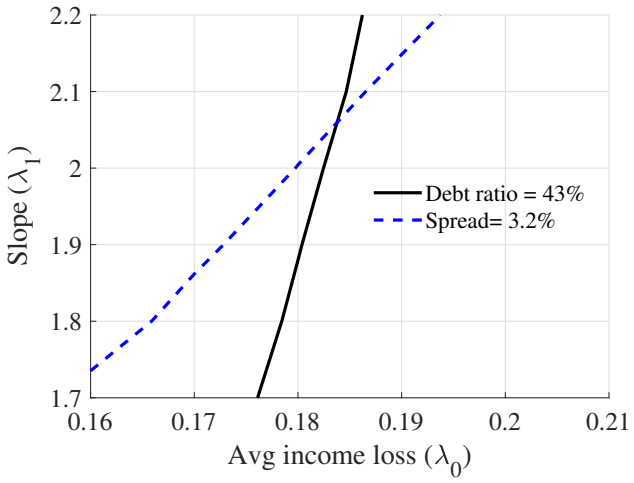


Figure 24.A.2: Combinations of parameters values (λ_0, λ_1) that match the target values for the debt-to-income ratio and spread.

The right panel of Figure 24.A.3 also illustrates how the value of the average cost parameter λ_0 has a significant effect on the debt level chosen by the sovereign in the simulations. For each value of λ_0, the panel presents the debt level chosen by the sovereign and the implied bond price when the sovereign enters the period with the mean debt level observed in the simulations. The panel shows that when λ_0 is higher, the sovereign chooses higher debt levels, exploiting the improved borrowing opportunities implied by the higher cost of defaulting.[46] In addition, the panel shows that the debt level chosen by the sovereign when λ_0 is higher implies a higher bond price. This is consistent with the lower spreads in the simulations for higher values of λ_0 presented in Figure 24.A.1.

Role of the parameter λ_1. Figure 24.A.4 illustrates how in comparison with the average cost parameter λ_0, the slope parameter λ_1 has a milder effect on the debt level chosen by the sovereign and a stronger effect on the spread paid by the sovereign. Everything else equal, an increase in the value of the slope parameter λ_1 makes defaulting less costly when income is below the mean and more costly when income is above the mean. This explains the shift of the income threshold at which the sovereign is indifferent between defaulting and repaying, depicted in the left panel of Figure 24.A.4. With a higher value of λ_1, for lower income levels, the cost of defaulting is lower and therefore, the income threshold (i.e., the maximum income for which the sovereign would choose to default) is higher. Consequently, with a higher value of λ_1, for lower debt levels, the set of income levels for which the sovereign would choose to default is larger (left panel of Figure 24.A.4). This implies that for these lower debt levels, lenders are willing to pay less for sovereign bonds, as illustrated in the right panel of Figure 24.A.4. In particular, the right panel of Figure 24.A.4 shows that an increase in the value of the slope parameter λ_1 lowers bond prices for the debt levels chosen by the sovereign in the

[46]The sovereign is eager to borrow more. In the benchmark calibration (of this and other quantitative studies of sovereign default), the sovereign is assumed to discount the future more than lenders. Therefore, without a borrowing constraint, the sovereign would choose debt levels significantly larger than the ones observed in the data. The inability to commit to repaying imposes an endogenous borrowing constraint on the sovereign. Increasing the cost of defaulting relaxes this constraint.

simulations. This is natural because defaults occur in periods of low income and, therefore, the cost of defaulting for low-income levels tends to be more relevant for bond prices.[47]

The right panel of Figure 24.A.4 also shows that equilibrium debt levels are very similar for different values of the slope parameter λ_1. However, with a higher value of λ_1, bond prices are significantly lower and promising to pay the same amount (i.e., the same debt level) allows the sovereign to borrow significantly less.

Overall, the equilibrium choices presented in the right panel of Figure 24.A.4 show that changing the value of the slope parameter λ_1 produces a large effect on the sovereign spread paid in equilibrium (and thus on the bond price) with a mild effect on the debt level. This contrasts with the effect on the equilibrium levels of debt and spread obtained by changing the value of the average cost parameter λ_0, which produces smaller changes in spread and larger changes in debt (right panel of Figure 24.A.3). This illustrates how the value of λ_1 is key for determining the spread level in the simulations, and the value of λ_0 is key for determining the debt level in the simulations, which is consistent with the existence of a unique combination of values of these parameters that allows us to match the calibration targets.

[47] The cost of defaulting for high-income levels only becomes relevant for very high debt levels, for which bond prices would be too low (right panel of Figure 24.A.4), implying spread levels inconsistent with those observed in the data.

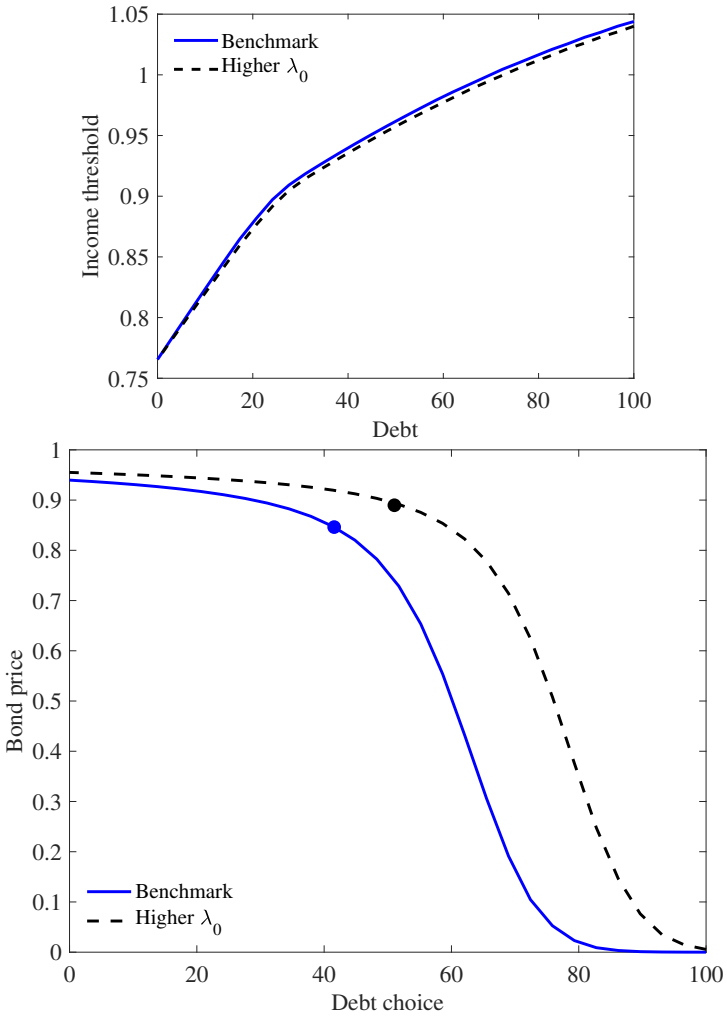


Figure 24.A.3: Default regions and bond price menus for different values of λ_0. Dashed lines assume $\lambda_0 = 0.19$ ($\lambda_0 = 0.184$ in the benchmark calibration). All other parameter values are the ones in the benchmark calibration. The left panel presents the income threshold at which the sovereign is indifferent between defaulting and repaying. The sovereign defaults for income levels below this threshold. The right panel assumes income is equal to the unconditional mean and presents the maximum price q lenders would be willing to pay for sovereign bonds. The solid dots in the right panel correspond to the optimal choices when the sovereign enters the period with the mean debt level observed in the simulations for each value of λ_0.

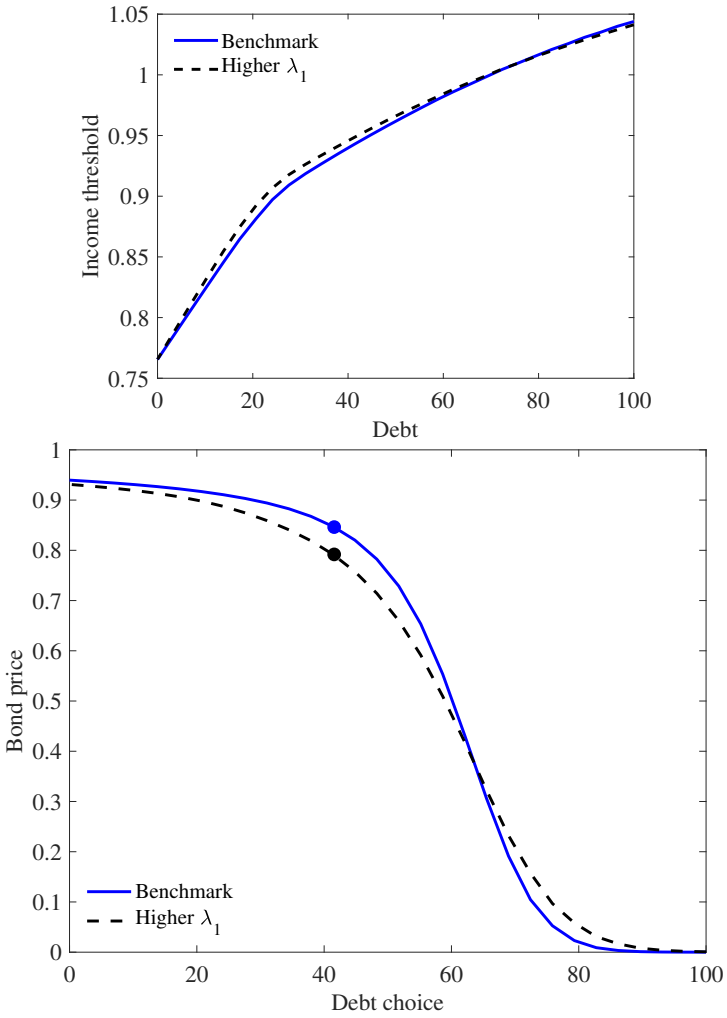


Figure 24.A.4: Default regions and bond price menus for different values of λ_1. Dashed lines assume $\lambda_1 = 2.5$ ($\lambda_1 = 2.059$ in the benchmark calibration). All other parameter values are the ones in the benchmark calibration. The left panel presents the income threshold at which the sovereign is indifferent between defaulting and repaying. The sovereign defaults for income levels below this threshold. The right panel assumes income is equal to the unconditional mean and presents the maximum price q lenders would be willing to pay for sovereign bonds. The solid dots in the right panel correspond to the optimal choices when the sovereign enters the period with the mean debt level observed in the simulations for each value of λ_1.

25.A Appendix to Chapter 25

One period is considered to be ten years. We set $\alpha = 0.3$ and $\nu = 0.05$ so that the shares of income that are allocated to capital and energy are 0.3 and 0.05, respectively. The discount factor is set to 0.985^{10} and the elasticity of substitution between the different energy inputs in (25.22) is set to 2 (which implies $\rho = 0.5$). In the same equation, λ_o, λ_c, and λ_g are set to 0.543, $0.187\lambda_o$, and $0.655\lambda_o$, respectively. TFP (i.e., z) is assumed to grow at 1.5%/year. K_0 and A_0 are set to 190 and 143, respectively. To calibrate the λs , prices and quantities of the three energy inputs are needed. Following Hassler et al. (2021a), we combine the demand equations (25.27) to derive the following relationships

$$\frac{\lambda_o}{\lambda_\kappa} = \left(\frac{e_{o,t}}{e_{\kappa,t}}\right)^{1-\rho} \frac{p_{1,t}}{p_{\kappa,t}},\ \kappa = c, g.$$

Using world market prices from Golosov et al. (2014), the coal price is set to (USD) \$74/ton. The oil price is calibrated to be about \$70/barrel. The relative price between oil and coal in units of carbon is then 5.87. Using the same source for the ratio of global oil to coal use in carbon units, we have $\lambda_o/\lambda_c = 5.348$. Turning to green energy, we employ data for the sum of nuclear, hydro, wind, waste, and other renewables from Golosov et al. (2014) and adopt their assumption of a unitary relative price between oil and green energy. This gives $\lambda_o/\lambda_g = 1.527$. Together with the normalization $\lambda_o + \lambda_c + \lambda_g = 1$, we arrive at the following values: $\lambda_o = 0.543$, $\lambda_c = 0.102$, and $\lambda_g = 0.356$.

Equation (25.10) is calibrated with $\phi_L = 0.2$, $\phi_0 = 0.393$, and $\phi = 0.0228$. The initial stock of carbon dioxide in the atmosphere, S_0, is set to 581 GtC. Finally, the parameters in equations (25.4)–(25.5) are calibrated as follows. Parameter $\sigma_1 = 0.22$, $\sigma_2 = 0.3$, $\sigma_3 = 0.05$, $\eta = 3.7$. The initial temperatures are set to $T_0 = 0.85$ and $T_{L,0} = 0.01$.