Part 1: Growth Theory (and introduction)

Chapters

1. Introduction 1
2. Growth Theory Preliminaries 7
3. The Neoclassical Growth Model 23
4. An Application: the Small Open Economy 41
5. Endogenous Growth I: Escaping Diminishing Returns 51
6. Endogenous Growth II: Technological Change 69
7. Proximate and Fundamental Causes of Growth 87

Filipe Campante,
Federico Sturzenegger
and Andrés Velasco

Advanced Macro-economics

LSE Press

Suggested citation: Campante, Filipe; Sturzenegger, Federico; and Velasco, Andrés. (2021) 'Introduction, and Part 1: Growth Theory', in Filipe Campante, Federico Sturzenegger and Andrés Velasco, *Advanced Macro-Economics: An Easy Guide*. London: LSE Press, 2021. Chapters 1–7. https://doi.org/10.31389/lsepress.ame.1

CHAPTER 1

Introduction

Paul Samuelson once stated that "macroeconomics, even with all of our computers and with all of our information is not an exact science and is incapable of being an exact science". Perhaps this quote captures the view that the field of macroeconomics, the study of aggregate behaviour of the economy, is full of loose ends and inconsistent statements that make it difficult for economists to agree on anything.

While there is truth to the fact that there are plenty of disagreements among macroeconomists, we believe such a negative view is unwarranted. Since the birth of macroeconomics as a discipline in the 1930s, in spite of all the uncertainties, inconsistencies, and crises, macroeconomic performance around the world has been strong. More recently, dramatic shocks, such as the Great Financial Crisis or the Covid pandemic, have been managed – not without cost, but with effective damage control. There is much to celebrate in the field of macroeconomics.

Macroeconomics was born under the pain of both U.S. and UK's protracted recession of the 1930s. Until then, economics had dealt with markets, efficiency, trade, and incentives, but it was never thought that there was place for a large and systematic breakdown of markets. High and persistent unemployment in the U.S. required a different approach.

The main distinctive feature to be explained was the large disequilibrium in the labour market. How could it be that a massive number of people wanted to work, but could not find a job? This led to the idea of the possibility of aggregate demand shortfalls – and thus of the potential role for government to prop it up, and, in doing so, restore economic normalcy. "Have people dig a hole and fill them up if necessary" is the oft-quoted phrase by Keynes. In modern economic jargon, increase aggregate demand to move the equilibrium of the economy to a higher level of output.

Thus, an active approach to fiscal and monetary policy developed, entrusting policy makers with the role of moderating the business cycle. The relationship was enshrined in the so-called Phillips curve, a relationship that suggested a stable tradeoff between output and inflation. If so, governments simply had to choose their preferred spot on that tradeoff.

Then things changed. Higher inflation in the 60s and 70s, challenged the view of a stable tradeoff between output and inflation. In fact, inflation increased with no gain in output, the age of stagflation had arrived. What had changed?

The answer had to do with the role of expectations in macroeconomics.[1]

The stable relationship between output and inflation required static expectations. People did not expect inflation, then the government found it was in its interest to generate a bit of inflation – but

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 1. 'Introduction', pp. 1–4. London: LSE Press. DOI: https://doi.org/10.31389/lsepress.ame.a

that meant people were always wrong! As they started anticipating the inflation, then its effect on employment faded away, and the effectiveness of macro policy had gone stale.

The rational expectations revolution in macroeconomics, initiated in the 1970s, imposed the constraint that a good macro model should allow agents in the model to understand it and act accordingly. This was not only a theoretical purism. It was needed to explain what was actually happening in the real world. The methodological change took hold very quickly and was embraced by the profession. As a working assumption, it is a ubiquitous feature of macroeconomics up to today.

Then an additional challenge to the world of active macroeconomic policy came about. In the early 1980s, some macroeconomists started the "real business cycles" approach: they studied the neoclassical growth model – that is, a model of optimal capital accumulation – but added to it occasional productivity shocks. The result was a simulated economy that, they argued, resembled on many dimensions the movements of the business cycle. This was a dramatic finding because it suggested that business cycles could actually be the result of optimal responses by rational economic agents, thereby eschewing the need for a stabilising policy response. What is more, active fiscal or monetary policy were not merely ineffective, as initially argued by the rational expectations view: they could actually be harmful.

This was the state of the discussion when a group of economists tackled the task of building a framework that recovered some of the features of the old Keynesian activism, but in a model with fully rational agents. They modelled price formation and introduced market structures that departed from a perfectly competitive allocation. They adhered strictly to the assumptions of rational expectations and optimisation, which had the added advantage of allowing for explicit welfare analyses. Thus, the New Keynesian approach was built. It also allowed for shocks, of course, and evolved into what is now known as dynamic stochastic general equilibrium (DSGE) models.

Macroeconomic policymaking evolved along those lines. Nowadays, DSGE models are used by any respectable central bank. Furthermore, because this type of model provides flexibility in the degree of price rigidities and market imperfections, it comprises a comprehensive framework nesting the different views about how individual markets operate, going all the way from the real business cycle approach to specifications with ample rigidities.

But the bottom line is that macroeconomics speaks with a common language. While differences in world views and policy preferences remain, having a common framework is a great achievement. It allows discussions to be framed around the parameters of a model (and whether they match the empirical evidence) – and such discussions can be more productive than those that swirl around the philosophical underpinnings of one's policy orientations.

This book, to a large extent, follows this script, covering the different views – and very importantly, the tools needed to speak the language of modern macroeconomic policymaking – in what we believe is an accessible manner. That language is that of dynamic policy problems.

We start with the Neoclassical Growth Model – a framework to think about capital accumulation through the lens of optimal consumption choices – which constitutes the basic grammar of that language of modern macroeconomics. It also allows us to spend the first half of the book studying economic growth – arguably the most important issue in macroeconomics, and one that, in recent decades, has taken up as much attention as the topic of business cycles. The study of growth will take us through the discussion of factor accumulation, productivity growth, the optimality of both the capital stock and the growth rate, and empirical work in trying to understand the proximate and fundamental causes of growth. In that process, we also develop a second canonical model in modern macroeconomics: the overlapping generations model. This lets us revisit some of the issues around capital accumulation and long-run growth, as well as study key policy issues, such as the design of pension systems.

We then move to discuss issues of consumption and investment. These are the key macroeconomic aggregates, of course, and their study allows us to explore the power of the dynamic tools we developed in the first part of the book. They also let us introduce the role of uncertainty and expectations, as well as the connections between macroeconomics and finance.

Then, in the second half of the book, we turn to the study of business cycle fluctuations, and what policy can and should do about it. We start with the real business cycle approach, as it is based on the neoclassical growth model. Then we turn to the Keynesian approach, starting from the basic IS-LM model, familiar to anyone with an undergraduate exposure to macroeconomics, but then showing how its modern version emerged: first, with the challenge of incorporating rational expectations, and then with the fundamentals of the New Keynesian approach. Only then, we present the canonical New Keynesian framework.

Once we've covered all this material, we discuss the scope and effectiveness of fiscal policy. We also go over what optimal fiscal policy would look like, as well as some of the reasons for why in practice it departs from those prescriptions. We then move to discuss monetary policy: the relationship between money and prices, the debate on rules vs discretion, and the consensus that arose prior to the 2008 financial crisis and the Great Recession. We then cover the post-crisis development of quantitative easing, as well as the constraints imposed by the zero lower bound on nominal interest rates. We finish off by discussing some current topics that have been influencing the thinking of policymakers on the fiscal and monetary dimensions: secular stagnation, the fiscal theory of the price level, and the role of asset-price bubbles and how policy should deal with them.

As you can see from this whirlwind tour, the book covers a lot of material. Yet, it has a clear methodological structure. We develop the basic tools in the first part of the book, making clear exactly what we need at each step. All you need is a basic knowledge of calculus, differential equations, and some linear algebra – and you can consult the mathematical appendix for the basics on the tools we introduce and use in the book. Throughout, we make sure to introduce the tools not for their own sake, but in the context of studying policy-relevant issues and helping develop a framework for thinking about dynamic policy problems. We then study a range of policy issues, using those tools to bring you to the forefront of macroeconomic policy discussions. At the very end, you will also find two appendices for those interested in tackling the challenge of running and simulating their own macroeconomic models.

All in all, Samuelson was right that macroeconomics cannot be an exact science. Still, there is a heck of a lot to learn, enjoy and discover – and this, we hope, will help you become an informed participant in exciting macroeconomic policy debates. Enjoy!

Note

[1] Surprisingly, the answer came from the most unexpected quarter: the study of agricultural markets. As early as 1960 John Muth was studying the cobweb model, a standard model in agricultural economics. In this model the farmers look at the harvest price to decide how much they plant, but then this provides a supply the following year which is inconsistent with this price. For example a bad harvest implies a high price, a high price implies lots of planting, a big harvest next year and thus a low price! The low price motivates less planting, but then the small harvest leads to a high price the following year! In this model, farmers were systematically wrong, and kept being wrong all the time. This is nonsense, argued Muth. Not only should they learn, they know the market and they should plant the equilibrium price, namely the price that induces the amount of planting that implies that next year that will be the price. There are no cycles, no mistakes, the market equilibrium holds from day one! Transferred to macroeconomic policy, something similar was happening.

Growth Theory

CHAPTER 2

Growth theory preliminaries

2.1 | Why do we care about growth?

It is hard to put it better than Nobel laureate Robert Lucas did as he mused on the importance of the study of economic growth for macroeconomists and for anyone interested in economic development.[1]

> 'The diversity across countries in measured per capita income levels is literally too great to be believed. (...) Rates of growth of real per capita GNP are also diverse, even over sustained periods. For 1960–80 we observe, for example: India, 1.4% per year; Egypt, 3.4%; South Korea, 7.0%; Japan, 7.1%; the United States, 2.3%; the industrial economies averaged 3.6%. (..) An Indian will, on average, be twice as well off as his grandfather; a Korean 32 times. (...) I do not see how one can look at figures like these without seeing them as representing possibilities. 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.*'
>
> Lucas Jr. (1988) (emphasis added)

While it is common to think about growth today as being somehow natural, even expected – in fact, if world growth falls from 3.5 to 3.2%, it is perceived as a big crisis – it is worthwhile to acknowledge that this was not always the case. Pretty much until the end of the 18th century growth was quite low, if it happened at all. In fact, it was so low that people could not see it during their lifetimes. They lived in the same world as their parents and grandparents. For many years it seemed that growth was actually *behind* as people contemplated the feats of antiquity without understanding how they could have been accomplished. Then, towards the turn of the 18th century, as shown in Figure 2.1 something happened that created explosive economic growth as the world had never seen before. Understanding this transition will be the purpose of Chapter 10. Since then, growth has become the norm. This is the reason the first half of this book, in fact up to Chapter 10, will deal with understanding growth. As we proceed we will ask about the determinants of capital accumulation (Chapters 2 through 5, as well as 8 and 9), and discuss the process of technological progress (Chapter 6). Institutional factors will be addressed in Chapter 7. The growth process raises many interesting questions: should we expect this growth to continue? Should we expect it eventually to decelerate? Or, on the contrary, will it accelerate without bound?

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 2. 'Growth theory preliminaries', pp. 7–22. London: LSE Press.
DOI: https://doi.org/10.31389/lsepress.ame.b

Figure 2.1 The evolution of the world GDP per capita over the years 1–2008

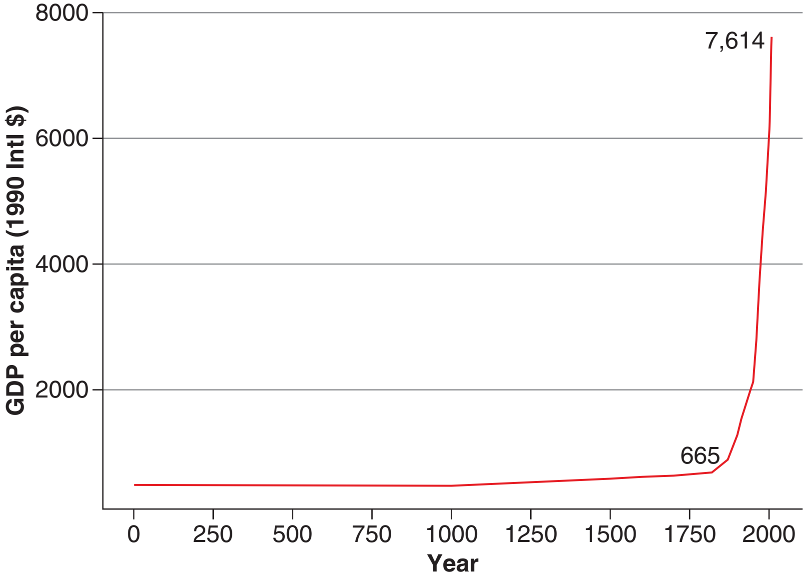
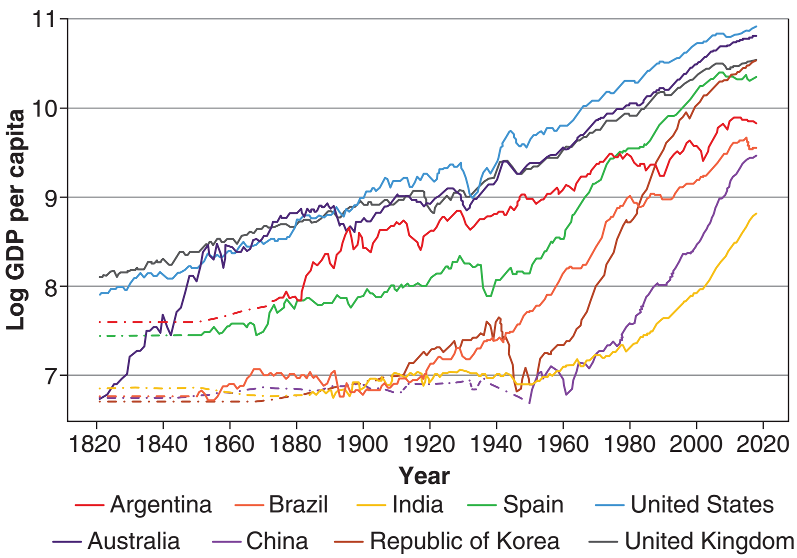
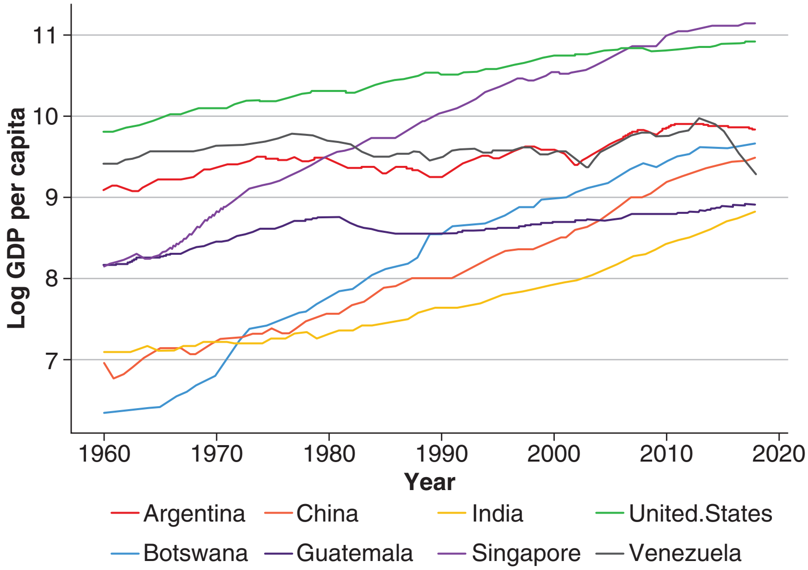


Figure 2.2 Log GDP per capita of selected countries (1820–2018)



But the fundamental point of Lucas's quote is to realise that the mind-boggling differences in income per capita across countries are to a large extent due to differences in growth rates over time; and the power of exponential growth means that even relatively small differences in the latter will build into huge differences in the former. Figures 2.2 and 2.3 make this point. The richest countries

Figure 2.3 Log GDP per capita of selected countries (1960–2018)



have been growing steadily over the last two centuries, and some countries have managed to converge to their income levels. Some of the performances are really stellar. Figure 2.2 shows how South Korea, with an income level that was 16% of that of the U.S. in 1940, managed to catch up in just a few decades. Today it's income is 68.5% compared to the U.S. Likewise, Spain's income in 1950 was 23% that of the U.S. Today it is 57%. At the same time other countries lagged. Argentina for example dropped from an income level that was 57% of U.S. income at the turn of the century to 33.5% today.

Figure 2.3 shows some diversity during recent times. The spectacular performances of Botswana, Singapore or, more recently, of China and India, contrast with the stagnation of Guatemala, Argentina or Venezuela. In 1960 the income of the average Motswana (as someone from Botswana is called) was only 6% as rich as the average Venezuelan. In 2018 he or she was 48% richer!

These are crucial reasons why we will spend about the initial half of this book in understanding growth. But those are not the only reasons! You may be aware that macroeconomists disagree on a lot of things; however, the issue of economic growth is one where there is much more of a consensus. It is thus helpful to start off on this relatively more solid footing. Even more importantly, the study of economic growth brings to the forefront two key ingredients of essentially all of macroeconomic analysis: general equilibrium and dynamics. First, understanding the behaviour of an entire economy requires thinking about how different markets interact and affect one another, which inevitably requires a general equilibrium approach. Second, to think seriously about how an economy evolves over time we must consider how today's choices affect tomorrow's – in other words, we must think dynamically! As such, economic growth is the perfect background upon which to develop the main methodological tools in macroeconomics: the model of intertemporal optimisation, known as the neoclassical growth model (NGM for short, also known as the Ramsey model), and the overlapping generations model (we'll call it the OLG model). A lot of what we will do later, as we explore different macroeconomic policy issues, will involve applications of these dynamic general-equilibrium tools that we will learn in the context of studying economic growth.

So, without further delay, to this we turn.

2.2 | The Kaldor facts

What are the key stylised facts about growth that our models should try to match? That there is growth in output and capital per worker with relatively stable income shares.

The modern study of economic growth starts in the post-war period and was mostly motivated by the experience of the developed world. In his classical article (Kaldor 1957), Nicolas Kaldor stated some basic facts that he observed economic growth seemed to satisfy, at least in those countries. These came to be known as the Kaldor facts, and the main challenge of growth theory as initially constituted was to account simultaneously for all these facts. But, what were these Kaldor facts? Here they are:[2]

1. *Output per worker shows continuous growth,* with no tendency to fall.
2. *The capital/output ratio is nearly constant.* (But what is capital?)
3. *Capital per worker shows continuous growth* (... follows from the other two).
4. *The rate of return on capital is nearly constant* (real interest rates are flat).
5. *Labour and capital receive constant shares of total income.*
6. *The growth rate of output per worker differs substantially across countries* (and over time, we can add, miracles and disasters).

Most of these facts have aged well. But not all of them. For example, we now know the constancy of the interest rate is not so when seen from a big historical sweep. In fact, interest rates have been on a secular downward trend that can be dated back to the 1300's (Schmelzing 2019). (Of course rates are way down now, so the question is how much lower can they go?) We will show you the data in a few pages.

In addition, in recent years, particularly since the early 1980s, the labour share has fallen significantly in most countries and industries. There is much argument in the literature as to the reasons why (see Karabarbounis and Neiman (2014) for a discussion on this) and the whole debate about income distribution trends recently spearheaded by Piketty (2014) has to do with this issue. We will come back to it shortly.

As it turns out Robert Solow established a simple model (Solow 1956) that became the first working model of economic growth.[3] Solow's contribution became the foundation of the NGM, and the backbone of modern growth theory, as it seemed to fit the Kaldor facts. Any study of growth must start with this model, reviewing what it explains – and, just as crucially, what it fails to explain.[4]

2.3 | The Solow model

We outline and solve the basic Solow model, introducing the key concepts of the neoclassical production function, the balanced growth path, transitional dynamics, dynamic inefficiency, and convergence.

Consider an economy with only two inputs: physical capital, K, and labour, L. The production function is

$$Y = F(K, L, t), \tag{2.1}$$

where Y is the flow of output produced. Assume output is a homogeneous good that can be consumed, C, or invested, I, to create new units of physical capital.

Let s be the fraction of output that is saved – that is, the *saving rate* – so that $1 - s$ is the fraction of output that is consumed. Note that $0 \leq s \leq 1$.

Assume that capital depreciates at the constant rate $\delta > 0$. The net increase in the stock of physical capital at a point in time equals gross investment less depreciation:

$$\dot{K} = I - \delta K = s \cdot F(K, L, t) - \delta K, \tag{2.2}$$

where a dot over a variable, such as $\dot{K}$, denotes differentiation with respect to time. Equation (2.2) determines the dynamics of K for a given technology and labour force.

Assume the population equals the labour force, L. It grows at a constant, exogenous rate, $\dot{L}/L = n \geq 0$.[5] If we normalise the number of people at time 0 to 1, then

$$L_t = e^{nt}. \tag{2.3}$$

where L_t is labour at time t.

If L_t is given from (2.3) and technological progress is absent, then (2.2) determines the time paths of capital, K, and output, Y. Such behaviour depends crucially on the properties of the production function, $F(\cdot)$. Apparently minor differences in assumptions about $F(\cdot)$ can generate radically different theories of economic growth.

2.3.1 | The (neoclassical) production function

For now, neglect technological progress. That is, assume that $F(\cdot)$ is independent of t. This assumption will be relaxed later. Then, the production function (2.1) takes the form

$$Y = F(K, L). \tag{2.4}$$

Assume also the following three properties are satisfied. First, for all $K > 0$ and $L > 0$, $F(\cdot)$ exhibits positive and diminishing marginal products with respect to each input:

$$\begin{aligned} \frac{\partial F}{\partial K} > 0, &\qquad \frac{\partial^2 F}{\partial K^2} < 0 \\ \frac{\partial F}{\partial L} > 0, &\qquad \frac{\partial^2 F}{\partial L^2} < 0. \end{aligned}$$

Second, $F(\cdot)$ exhibits constant returns to scale:

$$F(\lambda K, \lambda L) = \lambda \cdot F(K, L) \text{ for all } \lambda > 0.$$

Third, the marginal product of capital (or labour) approaches infinity as capital (or labour) goes to 0 and approaches 0 as capital (or labour) goes to infinity:

$$\begin{aligned} \lim_{K \to 0} \frac{\partial F}{\partial K} &= \lim_{L \to 0} \frac{\partial F}{\partial L} = \infty, \\ \lim_{K \to \infty} \frac{\partial F}{\partial K} &= \lim_{L \to \infty} \frac{\partial F}{\partial L} = 0. \end{aligned}$$

These last properties are called *Inada conditions*.

We will refer to production functions satisfying those three sets of conditions as *neoclassical production functions*.

The condition of constant returns to scale has the convenient property that output can be written as

$$Y = F(K, L) = L \cdot F(K/L, 1) = L \cdot f(k), \tag{2.5}$$

where $k \equiv K/L$ is the capital-labour ratio, and the function $f(k)$ is defined to equal $F(k, 1)$. The production function can be written as

$$y = f(k), \tag{2.6}$$

where $y \equiv Y/L$ is per capita output.

One simple production function that satisfies all of the above and is often thought to provide a reasonable description of actual economies is the Cobb-Douglas function,

$$Y = AK^{\alpha}L^{1-\alpha}, \tag{2.7}$$

where $A > 0$ is the level of the technology, and α is a constant with $0 < \alpha < 1$. The Cobb-Douglas function can be written as

$$y = Ak^{\alpha}. \tag{2.8}$$

Note that $f'(k) = A\alpha k^{\alpha-1} > 0, f''(k) = -A\alpha(1-\alpha)k^{\alpha-2} < 0, \lim_{k\to\infty} f'(k) = 0$, and $\lim_{k\to 0} f'(k) = \infty$. In short, the Cobb-Douglas specification satisfies the properties of a neoclassical production function.

2.3.2 | The law of motion of capital

The change in the capital stock over time is given by (2.2). If we divide both sides of this equation by L, then we get

$$\dot{K}/L = s \cdot f(k) - \delta k. \tag{2.9}$$

The right-hand side contains per capita variables only, but the left-hand side does not. We can write $\dot{K}/L$ as a function of k by using the fact that

$$\dot{k} \equiv \frac{d(K/L)}{dt} = \dot{K}/L - nk, \tag{2.10}$$

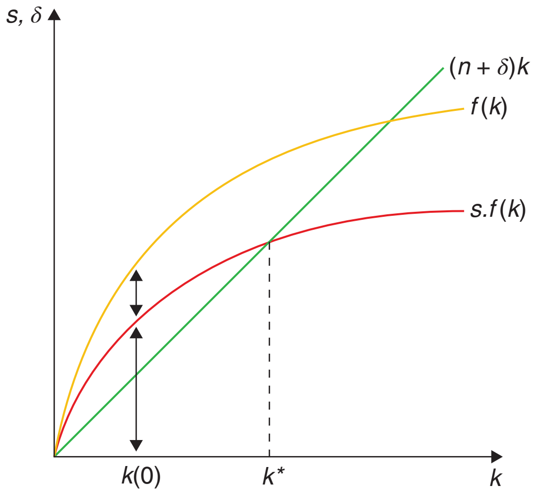
where $n = \dot{L}/L$. If we substitute (2.10) into the expression for $\dot{K}/L$ then we can rearrange terms to get

$$\dot{k} = s \cdot f(k) - (n + \delta) \cdot k. \tag{2.11}$$

The term $n + \delta$ on the right-hand side of (2.11) can be thought of as the effective depreciation rate for the capital/labour ratio, $k \equiv K/L$. If the saving rate, s, were 0, then k would decline partly due to depreciation of K at the rate δ and partly due to the growth of L at the rate n.

Figure 2.4 shows the workings of (2.11). The upper curve is the production function, $f(k)$. The term $s \cdot f(k)$ looks like the production function except for the multiplication by the positive fraction s. The $s \cdot f(k)$ curve starts from the origin (because $f(0) = 0$), has a positive slope (because $f'(k) > 0$), and gets flatter as k rises (because $f''(k) < 0$). The Inada conditions imply that the $s \cdot f(k)$ curve is vertical at $k = 0$ and becomes perfectly flat as k approaches infinity. The other term in (2.11), $(n+\delta) \cdot k$, appears in Figure 2.1 as a straight line from the origin with the positive slope $n + \delta$.

Figure 2.4 Dynamics in the Solow model



2.3.3 | Finding a balanced growth path

A *balanced growth path* (BGP) is a situation in which the various quantities grow at constant rates.[6] In the Solow model, the BGP corresponds to $\dot{k} = 0$ in (2.11).[7] We find it at the intersection of the $s \cdot f(k)$ curve with the $(n + \delta) \cdot k$ line in Figure 2.4. The corresponding value of k is denoted k^*. Algebraically, k^* satisfies the condition:

$$s \cdot f(k^*) = (n + \delta) \cdot k^*. \tag{2.12}$$

Since k is constant in the BGP, y and c are also constant at the values $y^* = f(k^*)$ and $c^* = (1 - s) \cdot f(k^*)$, respectively. Hence, in the Solow model, the per capita quantities k, y, and c do not grow in the BGP: it is a growth model without (long-term) growth!

Now, that's not quite right: the constancy of the per capita magnitudes means that the levels of variables – K, Y, and C – grow in the BGP at the rate of population growth, n. In addition, changes in the level of technology, represented by shifts of the production function, $f(\cdot)$; in the saving rate, s; in the rate of population growth, n; and in the depreciation rate, δ; all have effects on the per capita *levels* of the various quantities in the BGP.

We can illustrate the results for the case of a Cobb-Douglas production function. The capital/labour ratio on the BGP is determined from (2.12) as

$$k^* = \left(\frac{sA}{n + \delta}\right)^{\frac{1}{1-\alpha}}. \tag{2.13}$$

Note that, as we saw graphically for a more general production function $f(k)$, k^* rises with the saving rate, s, and the level of technology, A, and falls with the rate of population growth, n, and the depreciation rate, δ. Output per capita on the BGP is given by

$$y^* = A^{\frac{1}{1-\alpha}} \cdot \left(\frac{s}{n + \delta}\right)^{\frac{\alpha}{1-\alpha}}. \tag{2.14}$$

Thus, y^* is a positive function of s and A and a negative function of n and δ.

2.3.4 | Transitional dynamics

Moreover, the Solow model does generate growth in the transition to the BGP. To see the implications in this regard, note that dividing both sides of (2.11) by k implies that the growth rate of k is given by

$$\gamma_k \equiv \frac{\dot{k}}{k} = \frac{s \cdot f(k)}{k} - (n + \delta). \tag{2.15}$$

Equation (2.15) says that γ_k equals the difference between two terms, $s \cdot f(k)/k$ and $(n + \delta)$ which we plot against k in Figure 2.5. The first term is a downward-sloping curve, which asymptotes to infinity at $k = 0$ and approaches 0 as k tends to infinity. The second term is a horizontal line crossing the vertical axis at $n + \delta$. The vertical distance between the curve and the line equals the growth rate of capital per person, and the crossing point corresponds to the BGP. Since $n + \delta > 0$ and $s \cdot f(k)/k$ falls monotonically from infinity to 0, the curve and the line intersect once and only once. Hence (except for the trivial solution $k^* = 0$, where capital stays at zero forever), the BGP capital-labour ratio $k^* > 0$ exists and is unique.

Note also that output moves according to

$$\frac{\dot{y}}{y} = \alpha \frac{\dot{k}}{k} = \alpha \gamma_k. \tag{2.16}$$

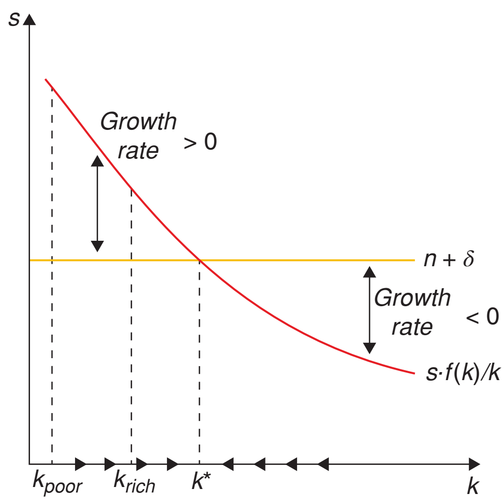
A formal treatment of dynamics follows. From (2.11) one can calculate

$$\frac{d\dot{k}}{dk} = s \cdot f'(k) - (n + \delta). \tag{2.17}$$

We want to study dynamics in the neighbourhood of the BGP, so we evaluate this at k^*:

$$\left.\frac{d\dot{k}}{dk}\right|_{k=k^*} = s \cdot f'(k^*) - (n + \delta). \tag{2.18}$$

Figure 2.5 Dynamics in the Solow model again



The capital stock will converge to its BGP if $\dot{k} > 0$ when $k < k^*$ and $\dot{k} < 0$ when $k > k^*$. Hence, this requires that $\left.\frac{d\dot{k}}{dk}\right|_{k=k^*} < 0$.

In the Cobb-Douglas case the condition is simple. Note that

$$\left.\frac{d\dot{k}}{dk}\right|_{k=k^*} = s \cdot A\alpha \left(\frac{sA}{n+\delta}\right)^{-1} - (n+\delta) = (n+\delta)(\alpha - 1) \tag{2.19}$$

so that $\left.\frac{d\dot{k}}{dk}\right|_{k=k^*} < 0$ requires $\alpha < 1$. That is, reaching the BGP requires diminishing returns.

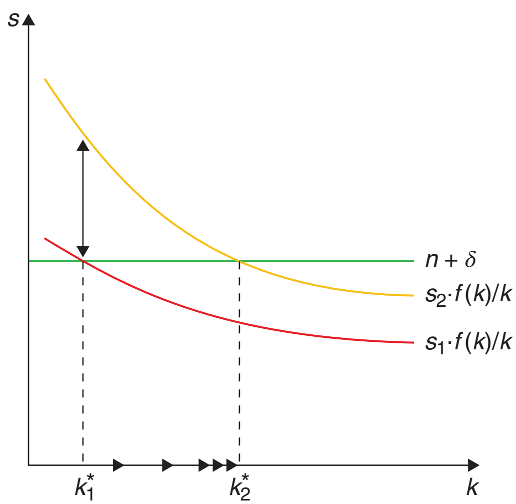
With diminishing returns, when k is relatively low, the marginal product of capital, $f'(k)$, is relatively high. By assumption, households save and invest a constant fraction, s, of this product. Hence, when k is relatively low, the marginal return to investment, $s \cdot f'(k)$, is relatively high. Capital per worker, k, effectively depreciates at the constant rate $n + \delta$. Consequently, the growth of capital, $\dot{k}$, is also relatively high. In fact, for $k < k^*$ it is positive. Conversely, for $k > k^*$ it is negative.

2.3.5 | Policy experiments

Suppose that the economy is initially on a BGP with capital per person k_1^*. Imagine that the government then introduces some policy that raises the saving rate permanently from s_1 to a higher value s_2. Figure 2.6 shows that the $s \cdot f(k)/k$ schedule shifts to the right. Hence, the intersection with the $n + \delta$ line also shifts to the right, and the new BGP capital stock, k_2^*, exceeds k_1^*. An increase in the saving rate generates temporarily positive per capita growth rates. In the long run, the levels of k and y are permanently higher, but the per capita growth rates return to 0.

A permanent improvement in the level of the technology has similar, temporary effects on the per capita growth rates. If the production function, $f(k)$, shifts upward in a proportional manner, then the

Figure 2.6 The effects of an increase in the savings rate



$s \cdot f(k)/k$ curve shifts upward, just as in Figure 2.6. Hence, γ_k again becomes positive temporarily. In the long run, the permanent improvement in technology generates higher levels of k and y, but no changes in the per capita growth rates.

2.3.6 | Dynamic inefficiency

For a given production function and given values of n and δ, there is a unique BGP value $k^* > 0$ for each value of the saving rate, s. Denote this relation by $k^*(s)$, with $dk^*(s)/ds > 0$. The level of per capita consumption on the BGP is $c^* = (1-s) \cdot f\left[k^*(s)\right]$. We know from (2.12) that $s \cdot f(k^*) = (n+\delta) \cdot k^*$; hence we can write an expression for c^*as

$$c^*(s) = f\left[k^*(s)\right] - (n+\delta) \cdot k^*. \tag{2.20}$$

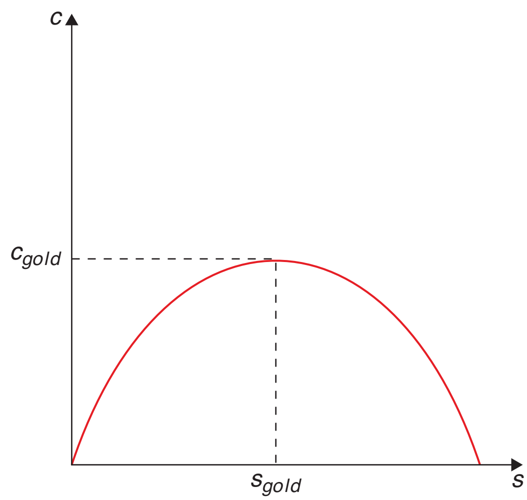
Figure 2.7 shows the relation between $c^*$and s that is implied by (2.20). The quantity $c^*$is increasing in s for low levels of s and decreasing in s for high values of s. The quantity c^* attains its maximum when the derivative vanishes, that is, when $\left[f'(k^*) - (n+\delta)\right] \cdot dk^*/ds = 0$. Since $dk^*/ds > 0$, the term in brackets must equal 0. If we denote the value of k^* by k_g that corresponds to the maximum of c^*, then the condition that determines k_g is

$$f'\left(k_g\right) = (n+\delta). \tag{2.21}$$

The corresponding savings rate can be denoted as s_g, and the associated level of per capita consumption on the BGP is given by $c_g = f\left(k_g\right) - (n+\delta) \cdot k_g$ and is is called the "golden rule" consumption rate.

If the savings rate is greater than that, then it is possible to increase consumption on the BGP, and also over the transition path. We refer to such a situation, where everyone could be made better off by an alternative allocation, as one of *dynamic inefficiency*. In this case, this dynamic inefficiency is brought about by oversaving: everyone could be made better off by choosing to save less and consume more. But this naturally begs the question: why would anyone pass up this opportunity? Shouldn't we

Figure 2.7 Feasible consumption



think of a better model of how people make their savings decisions? We will see about that in the next chapter.

2.3.7 | Absolute and conditional convergence

Equation (2.15) implies that the derivative of γ_k with respect to k is negative:

$$\partial\gamma_k/\partial k = \frac{s}{k}\left[f'(k) - \frac{f(k)}{k}\right] < 0. \tag{2.22}$$

Other things equal, smaller values of k are associated with larger values of γ_k. Does this result mean that economies with lower capital per person tend to grow faster in per capita terms? Is there *convergence* across economies?

We have seen above that economies that are structurally similar in the sense that they have the same values of the parameters s, n, and δ and also have the same production function, $F(\cdot)$, have the same BGP values k^* and y^*. Imagine that the only difference among the economies is the initial quantity of capital per person, $k(0)$. The model then implies that the less-advanced economies – with lower values of $k(0)$ and $y(0)$ – have higher growth rates of k. This hypothesis is known as *conditional convergence*: within a group of structurally similar economies (i.e. with similar values for s, n, and δ and production function, $F(\cdot)$), poorer economies will grow faster and catch up with the richer one. This hypothesis does seem to match the data – think about how poorer European countries have grown faster, or how the U.S. South has caught up with the North, over the second half of the 20th century.

An alternative, stronger hypothesis would posit simply that poorer countries would grow faster without conditioning on any other characteristics of the economies. This is referred to as *absolute convergence*, and does not seem to fit the data well.[8] Then again, the Solow model does *not* predict absolute convergence!

2.4 | Can the model account for income differentials?

We have seen that the Solow model does not have growth in per capita income in the long run. But can it help us understand income differentials?

We will tackle the empirical evidence on economic growth at a much greater level of detail later on. However, right now we can ask whether the simple Solow model can account for the differences in income levels that are observed in the world. According to the World Bank's calculations, the range of 2020 PPP income levels vary from \$ 138,000 per capita in Qatar or \$80,000 in Norway, all the way down to \$ 700 in Burundi. Can the basic Solow model explain this difference in income per capita of a factor of more than 100 times or even close to 200 times?

In order to tackle this question we start by remembering what output is supposed to be on the BGP:

$$y^* = A^{\frac{1}{1-\alpha}}\left(\frac{s}{n+\delta}\right)^{\frac{\alpha}{1-\alpha}}. \tag{2.23}$$

Assuming $A = 1$ and $n = 0$ this simplifies to:

$$y^* = \left(\frac{s}{\delta}\right)^{\frac{\alpha}{1-\alpha}}. \tag{2.24}$$

The ability of the Solow model to explain these large differences in income (in the BGP), as can be seen from the expressions above, will depend critically on the value of α.

$$\text{If}\begin{cases}\alpha = \frac{1}{3} \text{ then } \frac{\alpha}{1-\alpha} = \frac{1/3}{2/3} = \frac{1}{2} \\ \alpha = \frac{1}{2} \text{ then } \frac{\alpha}{1-\alpha} = \frac{1/2}{1/2} = 1 \\ \alpha = \frac{2}{3} \text{ then } \frac{\alpha}{1-\alpha} = \frac{2/3}{1/3} = 2. \end{cases}$$

The standard (rough) estimate for the capital share is $\frac{1}{3}$. Parente and Prescott (2002), however, claim that the capital share in GDP is much larger than usually accounted for because there are large intangible capital assets. In fact, they argue that the share of investment in GDP is closer to two-thirds rather than the more traditional one-third. The reasons for the unaccounted investment are (their estimates of the relevance of each in parenthesis):

1. Repair and maintenance (5% of GDP)
2. R&D (3% of GDP) multiplied by three (i.e. 9% of GDP) to take into account perfecting the manufacturing process and launching new products (the times three is not well substantiated)
3. Investment in software (3% of GDP)
4. Firms investment in organisation capital. (They think 12% is a good number.)
5. Learning on the job and training (10% of GDP)
6. Schooling (5% of GDP)

They claim all this capital has a return and that it accounts for about 56% of total GDP!

At any rate, using the equation above:

$$\frac{y_1}{y_2} = \frac{\left(\frac{s_1}{\delta}\right)^{\frac{\alpha}{1-\alpha}}}{\left(\frac{s_2}{\delta}\right)^{\frac{\alpha}{1-\alpha}}} = \left(\frac{s_1}{s_2}\right)^{\frac{\alpha}{1-\alpha}}, \tag{2.25}$$

which we can use to estimate income level differences.

	$\left(\frac{y_1}{y_2} - 1\right) * 100$		
$\frac{s_1}{s_2}$	$\alpha = \frac{1}{3}$	$\alpha = \frac{1}{2}$	$\alpha = \frac{2}{3}$
1	0%	0%	0%
1.5	22%	50%	125%
2	41%	100%	300%
3	73%	200%	800%

But even the 800% we get using the two-thirds estimate seems to be way too low relative to what we see in the data.

Alternatively, the differences in income may come from differences in total factor productivity (TFP), as captured by A. The question is: how large do these differences need to be to explain the output differentials? Recall from (2.23) that

$$y^* = A^{\frac{1}{1-\alpha}} \left(\frac{s}{n+\delta}\right)^{\frac{\alpha}{1-\alpha}}. \tag{2.26}$$

So if $\alpha = 2/3$, as suggested by Parente and Prescott (2002), then $A^{\frac{1}{1-\alpha}} = A^{\frac{1}{1/3}} = A^3$. Now, let's forget about s, δ, n (for example, by assuming they are the same for all countries), and just focus on differences in A. Notice that if TFP is $1/3$, of the level in the other country, this indicates that the income level is then $1/27$.

Parente and Prescott (2002) use this to estimate, for a group of countries, how much productivity would have to differ (relative to the United States) for us to replicate observed relative incomes over the period 1950–1988:

Country	Relative Income		Relative TFP
UK	60%	→	86%
Colombia	22%	→	64%
Paraguay	16%	→	59%
Pakistan	10%	→	51%

These numbers appear quite plausible, so the message is that the Solow model requires substantial cross-country differences in productivity to approximate existing cross-country differences in income. This begs the question of what makes productivity so different across countries, but we will come back to this later.

2.5 | The Solow model with exogenous technological change

We have seen that the Solow model does not have growth in per capita income in the long run. But that changes if we allow for technological change.

Allow now the productivity of factors to change over time. In the Cobb-Douglas case, this means that A increases over time. For simplicity, suppose that $\dot{A}/A = a > 0$. Out of the BGP, output then evolves according to

$$\frac{\dot{y}}{y} = \frac{\dot{A}}{A} + \alpha\frac{\dot{k}}{k} = a + \alpha\gamma_k. \tag{2.27}$$

On the BGP, where k is constant,

$$\frac{\dot{y}}{y} = a. \tag{2.28}$$

This is a strong prediction of the Solow model: in the long run, technological change is the only source of growth in per capita income.

Let's now embed this improvement in technology or efficiency in workers. We can define labour input as broader than just bodies, we could call it now human capital defined by

$$E_t = L_t \cdot e^{\lambda t} = L_0 \cdot e^{(\lambda+n)t}, \tag{2.29}$$

where E is the amount of labor in efficiency units. The production function is

$$Y = F\left(K_t, E_t\right). \tag{2.30}$$

To put it in per capita efficiency terms, we define

$$k = \frac{K}{E}. \tag{2.31}$$

So

$$\frac{\dot{k}}{k} = \frac{\dot{K}}{K} - \frac{\dot{E}}{E} = \frac{sy}{k} - \delta - n - \lambda, \tag{2.32}$$

$$\frac{\dot{k}}{k} = \frac{sf(k)}{k} - \delta - n - \lambda, \tag{2.33}$$

$$\dot{k} = sf(k) - (\delta + n + \lambda)\, k. \tag{2.34}$$

For $\dot{k} = 0$

$$\frac{sf(k)}{k} = (\delta + n + \lambda)\,. \tag{2.35}$$

On the BGP $\dot{k} = 0$, so

$$\frac{\dot{K}}{K} = \frac{\dot{E}}{E} = n + \lambda = \frac{\dot{Y}}{Y}. \tag{2.36}$$

But then

$$\frac{\dot{\left(\frac{Y}{L}\right)}}{\frac{Y}{L}} = \frac{\dot{Y}}{Y} - \frac{\dot{L}}{L} = \lambda \tag{2.37}$$

Notice that in this equilibrium income per person grows even on the BGP, and this accounts for all six Kaldor facts.

2.6 | What have we learned?

The Solow model shows that capital accumulation by itself cannot sustain growth in per capita income in the long run. This is because accumulation runs into diminishing marginal returns. At some point the capital stock becomes large enough – and its marginal product correspondingly small enough – that a given savings rate can only provide just enough new capital to replenish ongoing depreciation and increases in labour force. Alternatively, if we introduce exogenous technological change that increases productivity, we can generate long-run growth in income per capita, but we do not really explain it. In fact, any differences in long-term growth rates come from exogenous differences in the rate of technological change – we are not explaining those differences, we are just assuming them! As a result, nothing within the model tells you what policy can do about growth in the long run.

That said, we do learn a lot about growth in the transition to the long run, about differences in income levels, and how policy can affect those things. There are clear lessons about: (i) convergence – the model predicts conditional convergence; (ii) dynamic inefficiency – it is possible to save too much in this model; and (iii) long-run differences in income – they seem to have a lot to do with differences in productivity.

Very importantly, the model also points at the directions we can take to try and understand long-term growth. We can have a better model of savings behaviour: how do we know that individuals will save what the model says they will save? And, how does that relate to the issue of dynamic inefficiency?

We can look at different assumptions about technology: maybe we can escape the shackles of diminishing returns to accumulation? Or can we think more carefully about how technological progress comes about?

These are the issues that we will address over the next few chapters.

Notes

[1] Lucas's words hold up very well more than three decades later, in spite of some evidently dated examples.

[2] Once we are done with our study of economic growth, you can check the "new Kaldor facts" proposed by Jones and Romer (2010), which update the basic empirical regularities based on the progress over the subsequent half-century or so.

[3] For those of you who are into the history of economic thought, at the time the framework to study growth was the so-called Harrod-Domar model, due to the independent contributions of (you probably guessed it...) Harrod (1939) and Domar (1946). It assumed a production function with perfect complementarity between labour and capital ("Leontieff", as it is known to economists), and predicted that an economy would generate increasing unemployment of either labour or capital, depending on whether it saved a little or a lot. As it turns out, that was not a good description of the real world in the post-war period.

[4] Solow eventually got a Nobel prize for his trouble, in 1987 – also for his other contributions to the study of economic growth, to which we will return. An Australian economist, Trevor Swan, also published an independently developed paper with very similar ideas at about the same time, which is why sometimes the model is referred to as the Solow-Swan model. He did not get a Nobel prize.

[5] We will endogenise population growth in Chapter 10, when discussing unified growth theory.

[6] The BGP is often referred to as a "steady state", borrowing terminology from classical physics. We have noticed that talk of "steady state" tends to lead students to think of a situation where variables are not growing at all. The actual definition refers to constant growth rates, and it is only in certain cases and for certain variables, as we will see, that this constant rate happens to be zero.

[7] You should try to show mathematically from (2.11) that, with a neoclassical production function, the only way we can have a constant growth rate $\frac{\dot{k}}{k}$ is to have $\dot{k} = 0$.

[8] Or does it? More recently, Kremer et al. (2021) have argued that there has been a move towards absolute convergence in the data in the 21st century... Stay tuned!

References

Domar, E. D. (1946). Capital expansion, rate of growth, and employment. *Econometrica*, 137–147.

Harrod, R. F. (1939). An essay in dynamic theory. *The Economic Journal*, *49*(193), 14–33.

Jones, C. I. & Romer, P. M. (2010). The new Kaldor facts: Ideas, institutions, population, and human capital. *American Economic Journal: Macroeconomics*, *2*(1), 224–245.

Kaldor, N. (1957). A model of economic growth. *Economic Journal*, *67*(268), 591–624.

Karabarbounis, L. & Neiman, B. (2014). The global decline of the labor share. *The Quarterly Journal of Economics*, *129*(1), 61–103.

Kremer, M., Willis, J., & You, Y. (2021). Converging to convergence. *NBER Macro Annual 2021*. https://www.nber.org/system/files/chapters/c14560/c14560.pdf.

Lucas Jr., R. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, *22*, 3–42.

Parente, S. L. & Prescott, E. C. (2002). *Barriers to riches*. MIT Press.

Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press.

Schmelzing, P. (2019). Eight centuries of global real rates, r-g, and the 'suprasecular' decline, 1311–2018, Available at SSRN: https://ssrn.com/abstract=3485734 or http://dx.doi.org/10.2139/ssrn.3485734.

Solow, R. M. (1956). A contribution to the theory of economic growth. *The Quarterly Journal of Economics*, *70*(1), 65–94.

CHAPTER 3

The neoclassical growth model

3.1 | The Ramsey problem

We will solve the optimal savings problem underpinning the Neoclassical Growth Model, and in the process introduce the tools of dynamic optimisation we will use throughout the book. We will also encounter, for the first time, the most important equation in macroeconomics: the Euler equation.

$$\frac{\dot{c}_t}{c_t} = \sigma \left[f'\left(k_t\right) - \rho \right]$$

We have seen the lessons and shortcomings of the basic Solow model. One of its main assumptions, as you recall, was that the savings rate was constant. In fact, there was no optimisation involved in that model, and welfare statements are hard to make in that context. This is, however, a very rudimentary assumption for an able policy maker who is in possession of the tools of dynamic optimisation. Thus we tackle here the challenge of setting up an optimal program where savings is chosen to maximise intertemporal welfare.

As it turns out, British philosopher and mathematician Frank Ramsey, in one of the two seminal contributions he provided to economics before dying at the age of 26, solved this problem in 1928 (Ramsey (1928)).[1] The trouble is, he was so ahead of his time that economists would only catch up in the 1960s, when David Cass and Tjalling Koopmans independently revived Ramsey's contribution.[2] (That is why this model is often referred to either as the Ramsey model or the Ramsey-Cass-Koopmans model.) It has since become ubiquitous and, under the grand moniker of Neoclassical Growth Model (NGM), it is the foremost example of the type of *dynamic general equilibrium* model upon which the entire edifice of modern macroeconomics is built.

To make the problem manageable, we will assume that there is one representative household, all of whose members are both consumer and producer, living in a closed economy (we will lift this assumption in the next chapter). There is one good and no government. Each consumer in the representative household lives forever, and population growth is $n > 0$ as before. All quantities in small-case letters are per capita. Finally, we will look at the problem as solved by a benevolent central planner who maximises the welfare of that representative household, and evaluates the utility of future consumption at a discounted rate.

At this point, it is worth stopping and thinking about the model's assumptions. By now you are already used to outrageously unrealistic assumptions, but this may be a little too much. People

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 3. 'The neoclassical growth model', pp. 23–40. London: LSE Press.
DOI: https://doi.org/10.31389/lsepress.ame.c

obviously do not live forever, they are not identical, and what's this business of a benevolent central planner? Who are they? Why would they discount future consumption? Let us see why we use these shortcuts:

1. We will look at the central planner's problem, as opposed to the decentralised equilibrium, because it is easier and gets us directly to an efficient allocation. We will show that, under certain conditions, it provides the same results as the decentralised equilibrium. This is due to the so-called welfare theorems, which you have seen when studying microeconomics, but which we should perhaps briefly restate here:
   a. A competitive equilibrium is Pareto Optimal.
   b. All Pareto Optimal allocations can be decentralised as a competitive equilibrium under some convexity assumptions. Convexity of production sets means that we cannot have increasing returns to scale. (If we do, we need to depart from competitive markets.)
2. There's only one household? Certainly this is not very realistic, but it is okay if we think that typically people react similarly (not necessarily identically) to the parameters of the model. Specifically, do people respond similarly to an increase in the interest rate? If you think they do, then the assumption is okay.
3. Do all the people have the same utility function? Are they equal in all senses? Again, as above, not really. But, we believe they roughly respond similarly to basic tradeoffs. In addition, as shown by Caselli and Ventura (2000), one can incorporate a lot of sources of heterogeneity (namely, individuals can have different tastes, skills, initial wealth) and still end up with a representative household representation, as long as that heterogeneity has a certain structure. The assumption also means that we are, for the most part, ignoring distributional concerns, but that paper also shows that a wide range of distributional dynamics are compatible with that representation. (We will also raise some points about inequality as we go along.)
4. Do they live infinitely? Certainly not, but it does look like we have some intergenerational links. Barro (1974) suggests an individual who cares about the utility of their child: $u\left(c_t\right) + \beta V\left[u\left(c_{child}\right)\right]$. If that is the case, substituting recursively gives an intertemporal utility of the sort we have posited. And people do think about the future.
5. Why do we discount future utility? To some extent it is a revealed preference argument: interest rates are positive and this only makes sense if people value more today's consumption than tomorrow's, which is what we refer to when we speak of discounting the future. On this you may also want to check Caplin and Leahy (2004), who argue that a utility such as that in (3.1) imposes a sort of tyranny of the present: past utility carries no weight, whereas future utility is discounted. But does this make sense from a planner's point of view? Would this make sense from the perspective of tomorrow? In fact, Ramsey argued that it was unethical for a central planner to discount future utility.[3]

Having said that, let's go solve the problem.

3.1.1 | The consumer's problem

The utility function is[4]

$$\int_0^\infty u(c_t)e^{nt}e^{-\rho t}dt, \tag{3.1}$$

where c_t denotes consumption per capita and $\rho\ (> n)$ is the rate of time preference.[5] Assume $u'(c_t) > 0$, $u''(c_t) \leq 0$, and Inada conditions are satisfied.

3.1.2 | The resource constraint

The resource constraint of the economy is

$$\dot{K}_t = Y_t - C_t = F\left(K_t, L_t\right) - C_t, \tag{3.2}$$

with all variables as defined in the previous chapter. (Notice that for simplicity we assume there is no depreciation.) In particular, $F\left(K_t, L_t\right)$ is a neoclassical production function – hence neoclassical growth model. You can think of household production: household members own the capital and they work for themselves in producing output. Each member of the household inelastically supplies one unit of labour per unit of time.

This resource constraint is what makes the problem truly dynamic. The capital stock in the future depends on the choices that are made in the present. As such, the capital stock constitutes what we call the *state variable* in our problem: it describes the state of our dynamic system at any given point in time. The resource constraint is what we call the *equation of motion*: it characterises the evolution of the state variable over time. The other key variable, consumption, is what we call the *control variable*: it is the one variable that we can directly choose. Note that the control variable is jumpy: we can choose whatever (feasible) value for it at any given moment, so it can vary discontinuously. However, the state variable is sticky: we cannot change it discontinuously, but only in ways that are consistent with the equation of motion.

Given the assumption of constant returns to scale, we can express this constraint in per capita terms, which is more convenient. Dividing (3.2) through by L we get

$$\frac{\dot{K}_t}{L_t} = F\left(k_t, 1\right) - c_t = f\left(k_t\right) - c_t, \tag{3.3}$$

where $f(.)$ has the usual properties. Recall

$$\frac{\dot{K}_t}{L_t} = \dot{k}_t + nk_t. \tag{3.4}$$

Combining the last two equations yields

$$\dot{k}_t = f\left(k_t\right) - nk_t - c_t, \tag{3.5}$$

which we can think of as the relevant budget constraint. This is the final shape of the equation of motion of our dynamic problem, describing how the variable responsible for the dynamic nature of the problem – in this case the per capita capital stock k_t – evolves over time.

3.1.3 | Solution to consumer's problem

The household's problem is to maximise (3.1) subject to (3.5) for given k_0. If you look at our mathematical appendix, you will learn how to solve this, but it is instructive to walk through the steps here, as they have intuitive interpretations. You will need to set up the (current value) Hamiltonian for the problem, as follows:

$$H = u(c_t)e^{nt} + \lambda_t \left[f\left(k_t\right) - nk_t - c_t\right]. \tag{3.6}$$

Recall that c is the control variable (jumpy), and k is the state variable (sticky), but the Hamiltonian brings to the forefront another variable: λ, the *co-state variable*. It is the multiplier associated with the intertemporal budget constraint, analogously to the Lagrange multipliers of static optimisation.

Just like its Lagrange cousin, the co-state variable has an intuitive economic interpretation: it is the marginal value as of time t (i.e. the current value) of an additional unit of the state variable (capital, in this case). It is a (shadow) price, which is also jumpy.

First-order conditions (FOCs) are

$$\frac{\partial H}{\partial c_t} = 0 \Rightarrow u'(c_t)e^{nt} - \lambda_t = 0, \tag{3.7}$$

$$\dot{\lambda}_t = -\frac{\partial H}{\partial k_t} + \rho\lambda_t \Rightarrow \dot{\lambda}_t = -\lambda_t \left[f'\left(k_t\right) - n \right] + \rho\lambda_t, \tag{3.8}$$

$$\lim_{t\to\infty} \left(k_t \lambda_t e^{-\rho t} \right) = 0. \tag{3.9}$$

What do these optimality conditions mean? First, (3.7) should be familiar from static optimisation: differentiate with respect to the control variable, and set that equal to zero. It makes sure that, at any given point in time, the consumer is making the optimal decision – otherwise, she could obviously do better... The other two are the ones that bring the dynamic aspects of the problem to the forefront. Equation (3.9) is known as the transversality condition (TVC). It means, intuitively, that the consumer wants to set the optimal path for consumption such that, in the "end of times" (at infinity, in this case), they are left with no capital. (As long as capital has a positive value as given by λ, otherwise they don't really care...) If that weren't the case, I would be "dying" with valuable capital, which I could have used to consume a little more over my lifetime.

Equation (3.8) is the FOC with respect to the state variable, which essentially makes sure that at any given point in time the consumer is leaving the optimal amount of capital for the future. But how so? As it stands, it has been obtained mechanically. However, it is much nicer when we derive it purely from economic intuition. Note that we can rewrite it as follows:

$$\frac{\dot{\lambda}_t}{\lambda_t} = \rho - \left(f'\left(k_t\right) - n \right) \Rightarrow \rho + n = \frac{\dot{\lambda}_t}{\lambda_t} + f'\left(k_t\right). \tag{3.10}$$

This is nothing but an arbitrage equation for a typical asset price, where in this case the asset is the capital stock of the economy. Such arbitrage equations state that the opportunity cost of holding the asset (ρ in this case), equals its rate of return, which comprises the dividend yield ($f'(k_t) - n$) plus whatever capital gain you may get from holding the asset ($\frac{\dot{\lambda}_t}{\lambda_t}$). If the opportunity cost were higher (resp. lower), you would not be in an optimal position: you should hold less (resp. more) of the asset. We will come back to this intuition over and over again.

3.1.4 | The balanced growth path and the Euler equation

We are ultimately interested in the dynamic behaviour of our control and state variables, c_t and k_t. How can we turn our FOCs into a description of that behaviour (preferably one that we can represent graphically)? We start by taking (3.7) and differentiating both sides with respect to time:

$$u''(c_t)\dot{c}_t e^{nt} + nu'(c_t)e^{nt} = \dot{\lambda}_t. \tag{3.11}$$

Divide this by (3.7) and rearrange:

$$\frac{u''(c_t)c_t}{u'(c_t)}\frac{\dot{c}_t}{c_t} = \frac{\dot{\lambda}_t}{\lambda_t} - n. \tag{3.12}$$

Next, define

$$\sigma \equiv -\frac{u'(c_t)}{u''(c_t)c_t} > 0 \tag{3.13}$$

as the elasticity of intertemporal substitution in consumption.[6] Then, (3.12) becomes

$$\frac{\dot{c}_t}{c_t} = -\sigma\left(\frac{\dot{\lambda}_t}{\lambda_t} - n\right). \tag{3.14}$$

Finally, using (3.10) in (3.14) we obtain

$$\frac{\dot{c}_t}{c_t} = \sigma\left[f'\left(k_t\right) - \rho\right]. \tag{3.15}$$

This dynamic optimality condition is known as the Ramsey rule (or Keynes-Ramsey rule), and in a more general context it is referred to as the *Euler equation*. It may well be the most important equation in all of macroeconomics: it encapsulates the essence of the solution to any problem that trades off today versus tomorrow.[7]

But what does it mean intuitively? Think about it in these terms: if the consumer postpones the enjoyment of one unit of consumption to the next instant, it will be incorporated into the capital stock, and thus yield an extra $f'(\cdot)$. However, this will be worth less, by a factor of ρ. They will only consume more in the next instant (i.e. $\frac{\dot{c}_t}{c_t} > 0$) if the former compensates for the latter, as mediated by their proclivity to switch consumption over time, which is captured by the elasticity of intertemporal substitution, σ. Any dynamic problem we will see from now on involves some variation upon this general theme: the optimal growth rate trades off the rate of return of postponing consumption (i.e. investment) against the discount rate.

Mathematically speaking, equations (3.5) and (3.15) constitute a system of two differential equations in two unknowns. These plus the initial condition for capital and the TVC fully characterise the dynamics of the economy: once we have c_t and k_t, we can easily solve for any remaining variables of interest.

To make further progress, let us characterise the BGP of this economy. Setting (3.5) equal to zero yields

$$c^* = f(k^*) - nk^*, \tag{3.16}$$

which obviously is a hump-shaped function in c, k space. The dynamics of capital can be understood with reference to this function (Figure 3.1): for any given level of capital, if consumption is higher (resp. lower) than the BGP level, this means that the capital stock will decrease (resp. increase).

By contrast, setting (3.15) equal to zero yields

$$f'\left(k^*\right) = \rho. \tag{3.17}$$

This equation pins down the level of the capital stock on the BGP, and the dynamics of consumption can be seen in Figure 3.2: for any given level of consumption, if the capital stock is below (resp. above) its BGP level, then consumption is increasing (resp. decreasing). This is because the marginal product of capital will be relatively high (resp. low).

Expressions (3.16) and (3.17) together yield the values of consumption and the capital stock (both per-capita) in the BGP, as shown in Figure 3.3. This already lets us say something important about the behaviour of this economy. Let's recall the concept of the *golden rule*, from our discussion of the

Figure 3.1 Dynamics of capital

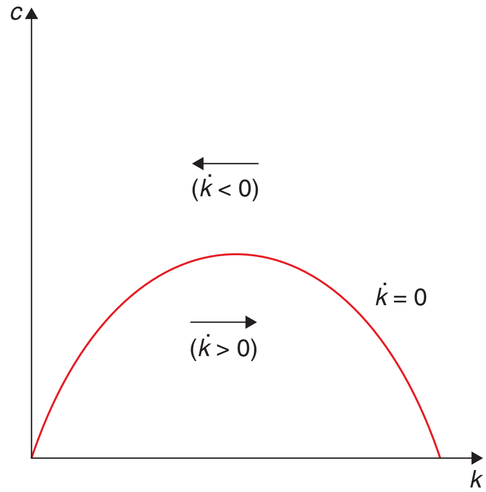
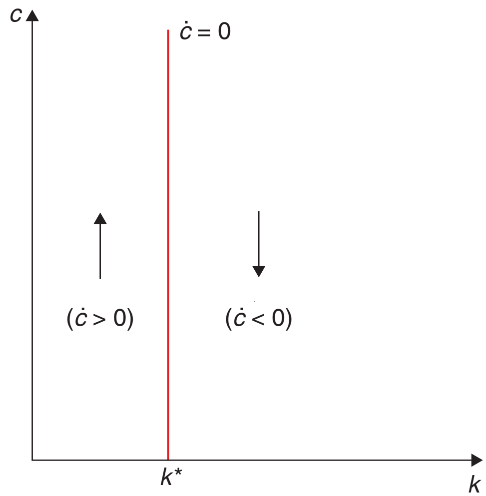


Figure 3.2 Dynamics of consumption



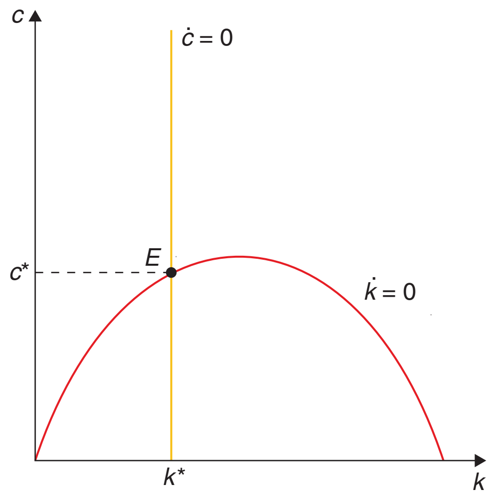
Solow model: the maximisation of per-capita consumption on the BGP. From (3.16) we see that this is tantamount to setting

$$\frac{\partial c^*}{\partial k^*} = f'\left(k^*_G\right) - n = 0 \Rightarrow f'\left(k^*_G\right) = n. \tag{3.18}$$

(Recall here we have assumed the depreciation rate is zero ($\delta = 0$).) If we compare this to (3.17), we see that the the optimal BGP level of capital per capita is *lower* than in the golden rule from the Solow model. (Recall the properties of the neoclassical production function, and that we assume $\rho > n$.)

Because of this comparison, (3.17) is sometimes known as the *modified golden rule*. Why does optimality require that consumption be lower on the BGP than what would be prescribed by the Solow

Figure 3.3 Steady state



golden rule? Because future consumption is discounted, it is not optimal to save so much that BGP consumption is maximised – it is best to consume more along the transition to the BGP. Keep in mind that it is (3.17), not (3.18), that describes the optimal allocation. The kind of oversaving that is possible in the Solow model disappears once we consider optimal savings decisions.

Now, you may ask: is it the case then that this type of oversaving is not an issue in practice (or even just in theory)? Well, we will return to this issue in Chapter 8. For now, we can see how the question of dynamic efficiency relates to issues of inequality.

3.1.5 | A digression on inequality: Is Piketty right?

It turns out that we can say something about inequality in the context of the NGM, even though the representative agent framework does not address it directly. Let's start by noticing that, as in the Solow model, on the BGP output grows at the rate n of population growth (since capital and output per worker are constant). In addition, once we solve for the decentralised equilibrium, which we sketch in Section 2 below, we will see that in that equilibrium we have $f'(k) = r$, where r is the interest rate, or equivalently, the rate of return on capital.

This means that the condition for dynamic efficiency, which holds in the NGM, can be interpreted as the $r > g$ condition made famous by Piketty (2014) in his influential *Capital in the 21st Century*. The condition $r > g$ is what Piketty calls the "Fundamental Force for Divergence": an interest rate that exceeds the growth rate of the economy. In short, he argues that, if $r > g$ holds, then there will be a tendency for inequality to explode as the returns to capital accumulate faster than overall income grows. In Piketty's words:

> 'This fundamental inequality (...) will play a crucial role in this book. In a sense, it sums up the overall logic of my conclusions. When the rate of return on capital significantly exceeds the growth rate of the economy (...), then it logically follows that inherited wealth grows faster than output and income.' (pp. 25–26)

Does that mean that, were we to explicitly consider inequality in a context akin to the NGM we would predict it to explode along the BGP? Not so fast. First of all, when taking the model to the data, we could ask what k is. In particular, k can have a lot of human capital i.e. be the return to labour mostly, and this may help undo the result. In fact, it could even turn it upside down if human capital is most of the capital and is evenly distributed in the population. You may also want to see Acemoglu and Robinson (2015), who have a thorough discussion of this prediction. In particular, they argue that, in a model with workers and capitalists, modest amounts of social mobility – understood as a probability that some capitalists may become workers, and vice-versa – will counteract that force for divergence.

Yet the issue has been such a hot topic in the policy debate that two more comments on this issue are due.

First, let's understand better the determinants of labour and income shares. Consider a typical Cobb-Douglas production function:

$$Y = AL^{\alpha}K^{1-\alpha}. \tag{3.19}$$

With competitive factor markets, the FOC for profit maximisation would give:

$$w = \alpha AL^{\alpha-1}K^{1-\alpha}. \tag{3.20}$$

Computing the labour share using the equilibrium wage gives:

$$\frac{wL}{Y} = \frac{\alpha AL^{\alpha-1}K^{1-\alpha}L}{AL^{\alpha}K^{1-\alpha}} = \alpha, \tag{3.21}$$

which implies that for a Cobb-Douglas specification, labour and capital shares are constant. More generally, if the production function is

$$Y = \left(\beta K^{\frac{\varepsilon-1}{\varepsilon}} + \alpha (AL)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}} \text{ with } \varepsilon \in [0, \infty), \tag{3.22}$$

then ε is the (constant) elasticity of substitution between physical capital and labour. Note that when $\varepsilon \to \infty$, the production function is linear (K and L are perfect substitutes), and one can show that when $\varepsilon \to 0$ the production function approaches the Leontief technology of fixed proportions, in which one factor cannot be substituted by the other at all.

From the FOC of profit maximisation we obtain:

$$w = \left(\beta K^{\frac{\varepsilon-1}{\varepsilon}} + \alpha (AL)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{1}{\varepsilon-1}} \alpha A (AL)^{-\frac{1}{\varepsilon}}, \tag{3.23}$$

the labour share is now:

$$\frac{wL}{Y} = \frac{\alpha\left(\beta K^{\frac{\varepsilon-1}{\varepsilon}} + \alpha (AL)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{1}{\varepsilon-1}} A^{\frac{\varepsilon-1}{\varepsilon}} L^{-\frac{1}{\varepsilon}} L}{\left(\beta K^{\frac{\varepsilon-1}{\varepsilon}} + \alpha (AL)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}}} = \alpha\left(\frac{AL}{Y}\right)^{\frac{\varepsilon-1}{\varepsilon}}. \tag{3.24}$$

Notice that as $\frac{L}{Y} \longrightarrow 0$, several things can happen to the labour share, and what happens depends on A and ε :

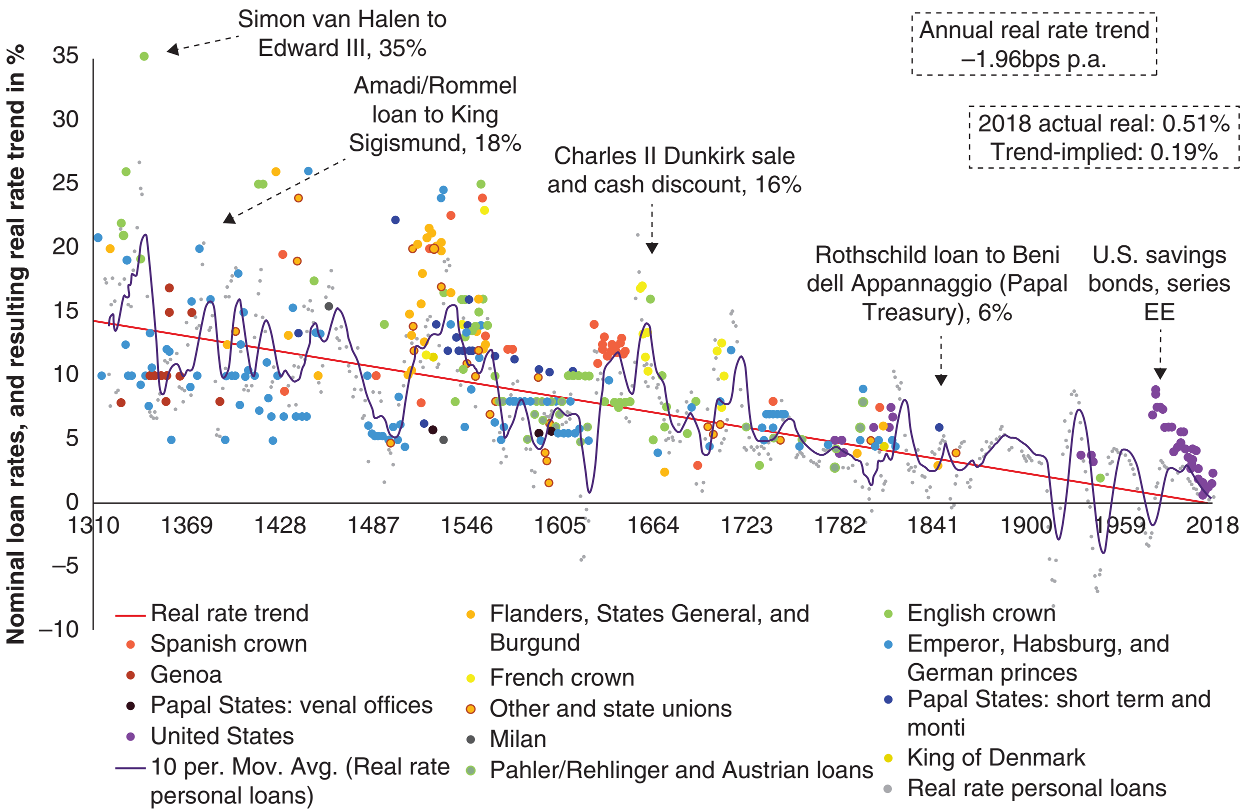
$$\text{If } \varepsilon > 1 \Longrightarrow \alpha\left(\frac{AL}{Y}\right)^{\frac{\varepsilon-1}{\varepsilon}} \longrightarrow 0 \tag{3.25}$$

$$\text{If } \varepsilon < 1 \Longrightarrow \alpha\left(\frac{AL}{Y}\right)^{\frac{\varepsilon-1}{\varepsilon}} \text{ increases.} \tag{3.26}$$

These two equations show that the elasticity of substitution is related to the concept of how essential a factor of production is. If the elasticity of substitution is less than one, the factor becomes more and more important with economic growth. If this factor is labour this may undo the Piketty result. This may be (and this is our last comment on the issue!) the reason why over the last centuries, while interest rates have been way above growth rates, inequality does not seem to have worsened. If anything, it seems to have moved in the opposite direction.

In Figure 3.4, Schmelzing (2019) looks at interest rates since the 1300s and shows that, while declining, they have consistently been above the growth rates of the economy at least until very recently. If those rates would have led to plutocracy, as Piketty fears, we would have seen it a long while ago. Yet the world seems to have moved in the opposite direction towards more democratic regimes.[8]

Figure 3.4 Real rates 1317–2018, from Schmelzing (2019)



3.1.6 | Transitional dynamics

How do we study the dynamics of this system? We will do so below graphically. But there are some shortcuts that allow you to understand the nature of the dynamic system, and particularly the relevant question of whether there is one, none, or multiple equilibria.

A dynamic system is a bunch of differential equations (difference equations if using discrete time). In the mathematical appendix, that you may want to refer to now, we argue that one way to approach this issue is to linearise the system around the steady state. For example, in our example here, Equations (3.5) and (3.15) are a system of two differential equations in two unknowns: c and k. To linearise the system around the BGP or steady state we compute the derivatives relative to each variable as shown below:

$$\begin{bmatrix} \dot{k}_t \\ \dot{c}_t \end{bmatrix} = \Omega \begin{bmatrix} k_t - k^* \\ c_t - c^* \end{bmatrix}, \tag{3.27}$$

where

$$\Omega = \begin{bmatrix} \left.\frac{\partial \dot{k}}{\partial k}\right|_{SS} & \left.\frac{\partial \dot{k}}{\partial c}\right|_{SS} \\ \left.\frac{\partial \dot{c}}{\partial k}\right|_{SS} & \left.\frac{\partial \dot{c}}{\partial c}\right|_{SS} \end{bmatrix} \tag{3.28}$$

and

$$\left.\frac{\partial \dot{k}}{\partial k}\right|_{SS} = f'(k^*) - n = \rho - n \tag{3.29}$$

$$\left.\frac{\partial \dot{k}}{\partial c}\right|_{SS} = -1 \tag{3.30}$$

$$\left.\frac{\partial \dot{c}}{\partial k}\right|_{SS} = \sigma c^* f''(k^*) \tag{3.31}$$

$$\left.\frac{\partial \dot{c}}{\partial c}\right|_{SS} = 0. \tag{3.32}$$

These computations allow us to define a matrix with the coefficients of the response of each variable to those in the system, at the steady state. In this case, this matrix is

$$\Omega = \begin{bmatrix} \rho - n & -1 \\ \sigma c^* f''(k^*) & 0 \end{bmatrix}. \tag{3.33}$$

In the mathematical appendix we provide some tools to understand the importance of this matrix of coefficients. In particular, this matrix has two associated eigenvalues, call them λ_1 and λ_2 (not to be confused with the marginal utility of consumption). The important thing to remember from the appendix is that the dynamic equations for the variables will be of the form $Ae^{\lambda_1} + Be^{\lambda_2}$. Thus, the nature of these eigenvalues turns out to be critical for understanding the dynamic properties of the system. If they are negative their effect dilutes over time (this configuration is called a sink, as variables converge to their steady state). If positive, the variable blows up (we call these systems a source, where variables drift away from the steady state). If one is positive and the other is negative the system typically blows up, except if the coefficient of the positive eigenvalue is zero (we call these saddle-path systems).

You may think that what you want is a sink, a system that converges to an equilibrium. While this may be the natural approach in sciences such as physics, this reasoning would not be correct in the realm of economics. Imagine you have one state variable (not jumpy) and a control variable (jumpy), as in this system. In the system we are analysing here k is a state variable that moves slowly over time and c

is the control variable that can jump. So, if you have a sink, you would find that *any c* would take you to the equilibrium. So rather than having a unique stable equilibrium you would have infinite alternative equilibria! Only if the two variables are state variables do you want a sink. In this case the equilibrium is unique because the state variables are uniquely determined at the start of the program.

In our case, to pin down a unique equilibria we would need a saddle-path configuration. Why? Because for this configuration there is only one value of the control variable that makes the coefficient of the explosive eigenvalue equal to zero. This feature is what allows to pin the unique converging equilibria. In the figures below this will become very clear.

What happens if all variables are control variables? Then you need the system to be a source, so that the control variables have only one possible value that attains sustainability. We will find many systems like this throughout the book.

In short, there is a rule that you may want to keep in mind. You need as many positive eigenvalues as jumpy or forward-looking variables you have in your system. If these two numbers match you have uniqueness![9]

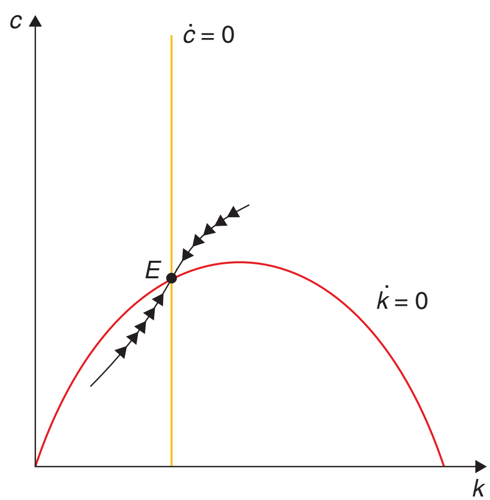
Before proceeding, one last rule you may want to remember. The determinant of the matrix is the product of the eigenvalues, and the trace is equal to the sum. This is useful, because, for example, in our two-equation model, if the determinant is negative, this means that the eigenvalues have different sign, indicating a saddle path. In fact, in our specific case,

- $\text{Det}(\Omega) = \sigma c^* f''(k^*) < 0$.

If $\text{Det}(\Omega)$ is the product of the eigenvalues of the matrix Ω and their product is negative, then we know that the eigenvalues must have the opposite sign. Hence, we conclude one eigenvalue is positive, while the other is negative.

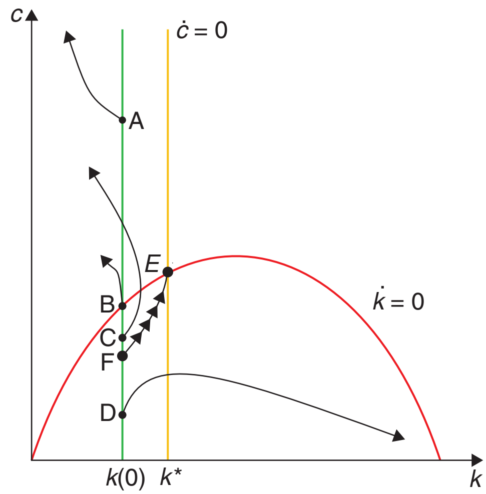
Recall that k is a slow-moving, or sticky, variable, while c can jump. Hence, since we have the same number of negative eigenvalues as of sticky variables, we conclude the system is saddle-path stable, and the convergence to the equilibrium unique. You can see this in a much less abstract way in the the phase diagram in Figure 3.5.

Figure 3.5 The phase diagram



Notice that since c can jump, from any initial condition for $k(0)$, the system moves vertically (c moves up or down) to hit the saddle path and converge to the BGP along the saddle path. Any other trajectory is divergent. Alternative trajectories appear in Figure 3.6.

Figure 3.6 Divergent trajectories



The problem is that these alternative trajectories either eventually imply a jump in the price of capital, which is inconsistent with rationality, or imply above-optimal levels of the capital stock. In either case this violates the transversality condition. In short, the first two dynamic equations provide the dynamics at any point in the (c,k) space, but only the TVC allows us to choose a single path that we will use to describe our equilibrium dynamics.[10]

3.1.7 | The effects of shocks

Consider the effects of the following shock. At time 0 and unexpectedly, the discount rate falls forever (people become less impatient). From the relevant $\dot{k} = 0$ and $\dot{c} = 0$ schedules, we see that the former does not move (ρ does not enter) but the latter does. Hence, the new BGP will have a higher capital stock. It will also have higher consumption, since capital and output are higher. Figure 3.7 shows the old BGP, the new BGP, and the path to get from one to the other. On impact, consumption falls (from point E to point A). Thereafter, both c and k rise together until reaching point E'.

Similar exercises can be carried out for other permanent and unanticipated shocks.

Consider, for example, an increase in the discount rate (Figure 3.8). (The increase is transitory, and that is anticipated by the planner.) The point we want to make is that there can be no anticipated jump in the control variables throughout the optimal path as this would allow for infinite capital gains. This is why the trajectory has to put you on the new saddle path when the discount rate goes back to normal.

Figure 3.7 A permanent increase in the discount rate

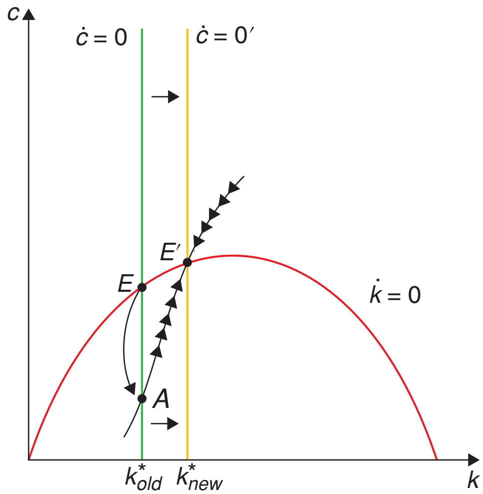
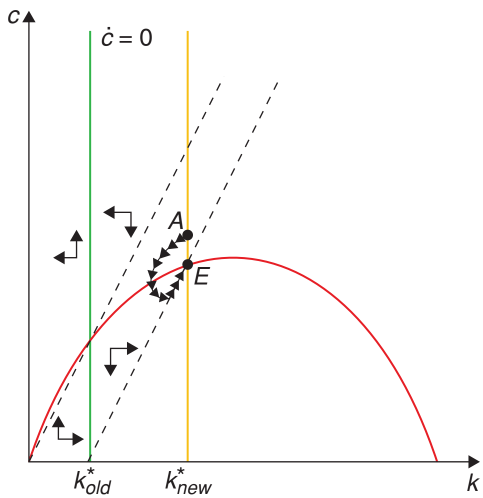


Figure 3.8 A transitory increase in the discount rate



3.2 | The equivalence with the decentralised equilibrium

We will show that the solution to the central planner's problem is exactly the same as the solution to a decentralised equilibrium.

Now we will sketch the solution to the problem of finding the equilibrium in an economy that is identical to the one we have been studying, but without a central planner. We now have households and firms (owned by households) who independently make their decisions in a perfectly competitive environment. We will only sketch this solution.

The utility function to be maximised by each household is

$$\int_0^\infty u(c_t)e^{nt}e^{-\rho t}dt, \tag{3.34}$$

where c_t is consumption and ρ ($> n$) is the rate of time preference.

The consumer's budget constraint can be written as

$$c_t L_t + \dot{A} = w_t L_t + rA_t, \tag{3.35}$$

where L_t is population, A_t is the stock of assets, $\dot{A}$ is the increase in assets, w_t is the wage per unit of labour (in this case per worker), and r is the return on assets. What are these assets? The households own the capital stock that they then rent out to firms in exchange for a payment of r; they can also borrow and lend money to each other, and we denote their total debt by B_t. In other words, we can define

$$A_t = K_t - B_t. \tag{3.36}$$

You should be able to go from (3.35) to the budget constraint in per worker terms:

$$c_t + \frac{da_t}{dt} + na_t = w_t + ra_t. \tag{3.37}$$

Households supply factors of production, and firms maximise profits. Thus, at each moment, you should be able to show that equilibrium in factor markets involves

$$r_t = f'\left(k_t\right), \tag{3.38}$$

$$w_t = f\left(k_t\right) - f'\left(k_t\right)k_t. \tag{3.39}$$

In this model, we must impose what we call a no-Ponzi-game (NPG) condition.[11] What does that mean? That means that households cannot pursue the easy path of getting arbitrarily rich by borrowing money and borrowing even more to pay for the interest owed on previously contracted debt. If possible that would be the optimal solution, and a rather trivial one at that. The idea is that the market will not allow these Ponzi schemes, so we impose this as a constraint on household behaviour.

$$\lim_{t\to\infty} a_t e^{-(r-n)t} \geq 0. \tag{3.40}$$

You will have noticed that this NPG looks a bit similar to the TVC we have seen in the context of the planner's problem, so it is easy to mix them up. Yet, they are different! The NPG is a *constraint* on optimisation – it wasn't needed in the planner's problem because there was no one in that closed economy from whom to borrow. In contrast, the TVC is an optimality condition – that is to say, something that is chosen in order to achieve optimality. They are related, in that both pertain to what happens in the limit, as $t \to \infty$. We will see how they help connect the decentralised equilibrium with the planner's problem.

3.2.1 | Integrating the budget constraint

The budget constraint in (3.37) holds at every instant t. It is interesting to figure out what it implies for the entire path to be chosen by households. To do this, we need to integrate that budget constraint. In future chapters we will assume that you know how to do this integration, and you can consult the mathematical appendix for that. But the first time we will go over all the steps.

So let's start again with the budget constraint for an individual family:

$$\dot{a}_t - (r - n)\, a_t = w_t - c_t. \tag{3.41}$$

This is a first-order differential equation which (as you can see in the Mathematical Appendix) can be solved using integrating factors. To see how that works, multiply both sides of this equation by $e^{-(r-n)t}$:

$$\dot{a}_t e^{-(r-n)t} + (n - r)\, a_t e^{-(r-n)t} = (w_t - c_t)e^{-(r-n)t}. \tag{3.42}$$

The left-hand side is clearly the derivative of $a_t e^{(n-r)t}$ with respect to time, so we can integrate both sides between 0 and t:

$$a_t e^{-(r-n)t} - a_0 = \int_0^t (w_s - c_s)e^{-(r-n)s} ds. \tag{3.43}$$

Taking the lim $t \longrightarrow \infty$ (and using the no-Ponzi condition) yields:

$$0 = \int_0^\infty \left(w_s - c_s\right) e^{-(r-n)s} ds + a_0, \tag{3.44}$$

which can be written as a standard intertemporal budget constraint:

$$\int_0^\infty w_s e^{-(r-n)s} ds + a_0 = \int_0^\infty c_s e^{-(r-n)s} ds. \tag{3.45}$$

This is quite natural and intuitive: all of what is consumed must be financed out of initial assets or wages (since we assume that Ponzi schemes are not possible).

3.2.2 | Back to our problem

Now we can go back to solve the consumer's problem

$$Max \int_0^\infty u(c_t)e^{nt}e^{-\rho t} dt \tag{3.46}$$

s.t.

$$c_t + \dot{a} + (n - r)\, a_t = w_t. \tag{3.47}$$

The Hamiltonian now looks like this

$$H = u\left(c_t\right) e^{nt} + \lambda_t \left[w_t - c_t - (n - r)\, a_t\right]. \tag{3.48}$$

From this you can obtain the FOCs and, following the same procedure from the previous case, you should be able to get to

$$-c_t \frac{u''\left(c_t\right)}{u'\left(c_t\right)} \frac{\dot{c}_t}{c_t} = (r - \rho). \tag{3.49}$$

How does that compare to (3.15), the Euler equation, which is one of our dynamic equations in the central planner's solution? We leave that to you.

You will also notice that, from the equivalent FOCs (3.7) and (3.8), we have

$$\frac{\dot{u}'}{u'} = (\rho - r), \tag{3.50}$$

or

$$u'\left(c_t\right) = e^{(\rho - r)t}. \tag{3.51}$$

Using this in the equivalent of (3.7) yields:

$$e^{(n-r)t} = \lambda_t e^{-\rho t}. \tag{3.52}$$

This means that the NPG becomes:

$$\lim_{t\to\infty} a_t \lambda_t e^{-\rho t} = \lim_{t\to\infty} a_t e^{-(r-n)t}. \tag{3.53}$$

You can show that this is exactly the same as the TVC for the central planner's problem. (Think about it: since all individuals are identical, what is the equilibrium level of b_t? If an individual wants to borrow, would anyone else want to lend?)

Finally, with the same reasoning on the equilibrium level of b_t, you can show that the resource constraint also matches the dynamic equation for capital, (3.5), which was the relevant resource constraint for the central planner's problem.

3.3 | Do we have growth after all?

Not really.

Having seen the workings of the Ramsey model, we can see that on the BGP, just as in the Solow model, there is no growth in per capita variables: k is constant at k^* such that $f'(k^*) = \rho$, and y is constant at $f(k^*)$. (It is easy to show that once again we can obtain growth if we introduce exogenous technological progress.)

3.4 | What have we learned?

We are still left with a growth model without long-run growth: it was not the exogeneity of the savings rate that generated the unsatisfactory features of the Solow model when it comes to explaining long-run growth. We will have to keep looking by moving away from diminishing returns or by modelling technological progress.

On the other hand, our exploration of the Ramsey model has left us with a microfounded framework that is the foundation of a lot of modern macroeconomics. This is true not only of our further explorations that will lead us into endogenous growth, but eventually also when we move to the realm of short term fluctuations. At any rate, the NGM is a dynamic general equilibrium framework that we will use over and over again.

Even in this basic application some key results have emerged. First, we have the Euler equation that encapsulates how consumers make optimal choices between today and tomorrow. If the marginal benefit of reducing consumption – namely, the rate of return on the extra capital you accumulate –

is greater than the consumer's impatience – the discount rate – then it makes sense to postpone consumption. This crucial piece of intuition will appear again and again as we go along in the book, and is perhaps the key result in modern macroeconomics. Second, in this context there is no dynamic inefficiency, as forward-looking consumers would never choose to oversave in an inefficient way.

Most importantly, now we are in possession of a powerful toolkit for dynamic analysis, and we will make sure to put it to use from now on.

Notes

1 The other one was to the theory of optimal taxation (Ramsey 1927).

2 See Cass (1965) and Koopmans et al. (1963).

3 Another interesting set of questions refer to population policies: say you impose a policy to reduce population growth. How does that play into the utility function? How do you count people that have not been and will not be born? Should a central planner count those people?

4 We are departing from standard mathematical convention, by using subscripts instead of parentheses to denote time, even though we are modelling time as continuous and not discrete. We think it pays off to be unconventional, in terms of making notation look less cluttered, but we apologise to the purists in the audience nonetheless!

5 Note that we must assume that $\rho > n$, or the problem will not be well-defined. Why? Because if $\rho < n$, the representative household gets more total utility out of a given level of consumption per capita in the future as there will be more "capitas" then. If the discount factor does not compensate for that, it would make sense to always postpone consumption! And why do we have e^{nt} in the utility function in the first place? Because we are incorporating the utility of all the individuals who are alive at time t – the more, the merrier!

6 Recall that the elasticity of a variable x with respect to another variable y is defined as $\frac{\frac{dx}{dy}}{\frac{x}{y}}$. As such, $\frac{1}{\sigma}$ is the elasticity of the marginal utility of consumption with respect to consumption – it measures how sensitive the marginal utility is to increases in consumption. Now, think about it: the more sensitive it is, the more I will want to smooth consumption over time, and this means I will be less likely to substitute consumption over time. That is why the inverse of that captures the intertemporal elasticity of substitution: the greater σ is, the more I am willing to substitute consumption over time.

7 This is the continuous-time analogue of the standard optimality condition that you may have encountered in microeconomics: the marginal rate of substitution (between consumption at two adjacent points in time) must be equal to the marginal rate of transformation.

8 At any rate, it may also be argued that maybe we haven't seen plutocracies because Piketty was right. After all, the French and U.S. revolutions may be explained by previous increases in inequality.

9 It works the same for a system of difference equation in discrete time, except that the cutoff is with eigenvalues being larger or smaller than one.

10 To rule out the path that leads to the capital stock of when the $\dot{k} = 0$ locus crosses the horizontal axis to the right of the golden rule, notice that λ from (3.8) grows at the rate $\rho + n - f'(k)$ so that $\lambda e^{-\rho t}$ grows at rate $n - f'(k)$, but to the right of the golden rule $n > f'(k)$, so that the term increases. Given that the capital stock is eventually fixed we conclude that the transversality condition cannot hold. The paths that lead to high consumption and a zero capital stock imply a collapse of consumption to zero when the path reaches the vertical axis. This trajectory is not feasible because at some point

it cannot continue. When that happens the price of capital increases, and consumers would have arbitraged that jump away, so that that path would have not occurred in the first place.

[11] Or should it now be the no-Madoff-game condition?

References

Acemoglu, D. & Robinson, J. A. (2015). The rise and decline of general laws of capitalism. *Journal of Economic Perspectives*, *29*(1), 3–28.

Barro, R. J. (1974). Are government bonds net wealth? *Journal of Political Economy*, *82*(6), 1095–1117.

Caplin, A. & Leahy, J. (2004). The social discount rate. *Journal of Political Economy*, *112*(6), 1257–1268.

Caselli, F. & Ventura, J. (2000). A representative consumer theory of distribution. *American Economic Review*, *90*(4), 909–926.

Cass, D. (1965). Optimum growth in an aggregative model of capital accumulation. *The Review of Economic Studies*, *32*(3), 233–240.

Koopmans, T. C. et al. (1963). *On the concept of optimal economic growth* (tech. rep.) Cowles Foundation for Research in Economics, Yale University.

Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press.

Ramsey, F. P. (1927). A contribution to the theory of taxation. *The Economic Journal*, *37*(145), 47–61.

Ramsey, F. P. (1928). A mathematical theory of saving. *The Economic Journal*, *38*(152), 543–559.

Schmelzing, P. (2019). Eight centuries of global real rates, r-g, and the 'suprasecular' decline, 1311–2018, Available at SSRN: https://ssrn.com/abstract=3485734 or http://dx.doi.org/10.2139/ssrn.3485734.

CHAPTER 4

An application: The small open economy

The Neoclassical Growth Model (NGM) is more than just a growth model. It provides us with a powerful tool to think about a number of questions in macroeconomics, which require us to think dynamically. So let's now put it to work!

We will do so in a simple but important application: understanding the capital accumulation dynamics for a small open economy. As we will see shortly, in an open economy the capital accumulation process is modified by the possibility of using foreign savings. This really allows countries to move much faster in the process of capital accumulation (if in doubt ask the Norwegians!), and is one of the main reasons why integrating into world capital markets is often seen as a big positive for any economy.

The use of foreign savings (or the accumulation of assets abroad) is summarised in the economy's current account, so the NGM applied to a small open economy can be thought of as yielding a model of the behaviour of the current account. The current account provides a measure of how fast the country is building foreign assets (or foreign debt), and as such it is a key piece to assess the sustainability of macroeconomic policies. We will also see that the adjustment of the current account to different shocks can lead to surprising and unexpected results. Finally, the framework can be used to analyse, for example, the role of stabilisation funds in small open economies.

4.1 | Some basic macroeconomic identities

A quick refresher that introduces the concept of the current account.

A good starting point is to start with the basic macroeconomic identities, which you have seen before in your introductory and intermediate macro courses. Recall the relationship between GNP (Gross National Product, the amount that is paid to a country's residents) and GDP (Gross Domestic Product, the amount of final goods produced in a particular country):

$$GDP + rB = GNP, \tag{4.1}$$

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 4. 'An application: The small open economy', pp. 41–50. London: LSE Press.
DOI: https://doi.org/10.31389/lsepress.ame.d

where B is the position held by residents in foreign assets (net of domestic assets held by foreigners), and r is the rate of return paid on those assets. In other words, rB is the total (net) interest payments made by foreigners to residents. Notice that countries with foreign assets will have a GNP which is larger than GDP, whereas countries with debt will have a GNP which is smaller than their GDP.

Starting from the output equation we have that

$$GNP = C + I + G + X - M + rB, \tag{4.2}$$

where C, I, G, X and M, stand, as usual, for consumption, investment, government expenditures, exports and imports. This can be rewritten as:

$$\underbrace{GNP - C - G}_{} - I = X - M + rB = CA, \tag{4.3}$$

$$S - I = X - M + rB = CA. \tag{4.4}$$

The right-hand side (RHS) is roughly the current account CA (the trade balance plus the net income on net foreign assets, which is typically called primary income, add up to the current account).[1] The equation says that the current account is equal to the difference between what the economy saves (S) and what it invests (I).[2]

Another alternative is to write this as:

$$GNP - \underbrace{C - G - I}_{} = X - M + rb = CA,$$

$$Y - Absorption = X - M + rb = CA,$$

which clearly shows that a current account is the difference between income and absorption. In common parlance: if a country spends more than it earns, it runs a current account deficit. Importantly, and as we will see over and over again, this does not mean that current account deficits are bad! They simply mean that a country is using debt and, as always, whether that is a good or a bad thing hinges on whether that is done in a sustainable way. To figure that out, we need to embed these accounting identities into an optimising intertemporal model of how consumption and investment are chosen given current and future output.

As luck would have it, this is exactly what the NGM is!

4.2 | The Ramsey problem for a small open economy

We will solve the (benevolent central planner) Ramsey problem for a small open economy. The key conclusions are: (i) $c = c^*$: consumption can be perfectly smoothed; (ii) $f'(k^*) = r$: the capital stock can adjust immediately via foreign borrowing, and thus becomes independent of domestic savings. This is because the current account allows the economy to adjust to shocks while maintaining optimal levels of consumption and capital stock.

Here is where we will start using, right away, what we learnt in the previous chapter. As before, there is one infinitely-lived representative household whose members consume and produce. Population growth is now assumed to be $n = 0$ for simplicity; initial population L_0 is normalised to 1, so that all quantities are in per capita terms (in small-case letters). There is one good, and no government.

The key difference is that now the economy is open, in the sense that foreigners can buy domestic output, and domestic residents can buy foreign output. Whenever domestic income exceeds domestic expenditure, locals accumulate claims on the rest of the world, and vice versa. These claims take the form of an internationally-traded bond, denominated in units of the only good. The economy is also small, in the sense that it does not affect world prices (namely the interest rate), and thus takes them as given.

We will assume also that the country faces a constant interest rate. The constancy of r is a key defining feature of the small country model. However, this is a strong assumption – if a country borrows a lot, probably its country risk would increase and so would the interest rate it would have to pay – but we will keep this assumption for now. (We will return to issues of risk when we talk about consumption and investment, in (Chapters 13 and 14.)

The utility function is exactly as before (with $n = 0$):

$$\int_0^\infty u(c_t)e^{-\rho t}dt. \tag{4.5}$$

The resource constraint of the economy is

$$\dot{k}_t + \dot{b}_t = f\left(k_t\right) + rb_t - c_t. \tag{4.6}$$

The novelty is that now domestic residents own a stock b_t of the bond, whose rate of return is r, which is a constant from the standpoint of the small open economy. What is the current account in this representation? It is income (GNP), which is $f\left(k_t\right) + rb_t$, minus consumption c_t, minus investment $\dot{k}_t$. In other words, it is equal to $\dot{b}_t$. A current-account surplus is equivalent to an increase in foreign bond holdings.

In the open economy, we also have to impose a no-Ponzi game (NPG) condition (or solvency condition):

$$\lim_{T\to\infty}\left(B_T e^{-rT}\right) = \lim_{T\to\infty}\left(b_T e^{-rT}\right) \geq 0. \tag{4.7}$$

This condition – again, not to be confused with the transversality condition (TVC) we met in the previous chapter – did not apply to the benevolent central planner (BCP) in the last chapter because they could not borrow in the context of a closed economy. It did apply to the consumers in the decentralised equilibrium though, and here it must apply to the economy as a whole. It means that the economy cannot run a Ponzi scheme with the rest of the world by borrowing money indefinitely to pay interest on its outstanding debt. In other words, this rules out explosive trajectories of debt accumulation under the assumption that foreigners would eventually stop lending money to this pyramid scheme.

4.2.1 | A useful transformation

Define total domestic assets per capita as

$$a_t = k_t + b_t. \tag{4.8}$$

Then, (4.6) becomes

$$\dot{a}_t = ra_t + f\left(k_t\right) - rk_t - c_t, \tag{4.9}$$

and (4.7) becomes

$$\lim_{T\to\infty}\left[\left(a_T - k_T\right)e^{-rT}\right] \geq 0. \tag{4.10}$$

4.2.2 | Solution to consumer's problem

The consumer maximises (4.5) subject to (4.9) and (4.10) for given k_0 and b_0. The Hamiltonian for the problem can be written as

$$H = u(c_t) + \lambda_t \left[ra_t + f\left(k_t\right) - rk_t - c_t\right]. \tag{4.11}$$

Note c is one control variable (jumpy), and k now is another control variable (also jumpy). It is now a that is the state variable (sticky), the one that has to follow the equation of motion. λ is the costate variable (the multiplier associated with the intertemporal budget constraint, also jumpy). The costate has the same intuitive interpretation as before: the marginal value as of time t of an additional unit of the state (assets a, in this case). (Here is a question for you to think about: why is capital a jumpy variable now, while it used to be sticky in the closed economy?)

The first order conditions are then:

$$u'\left(c_t\right) = \lambda_t, \tag{4.12}$$

$$f'\left(k_t\right) = r, \tag{4.13}$$

$$\dot{\lambda} = -r\lambda_t + \rho\lambda_t, \tag{4.14}$$

and

$$\lim_{t\to\infty} a_t\lambda_t e^{-\rho t} = 0. \tag{4.15}$$

Using (4.12) in (4.14), we obtain

$$u''\left(c_t\right)\dot{c}_t = (-r+\rho)u'\left(c_t\right). \tag{4.16}$$

Dividing both sides by $u'\left(c_t\right)$ and using the definition of the elasticity of intertemporal substitution, σ, gets us to our Euler equation for the dynamic behaviour of consumption:

$$\frac{\dot{c}_t}{c_t} = \sigma\left(r-\rho\right). \tag{4.17}$$

This equation says that per-capita consumption is constant only if $r = \rho$, which we assume from now on. Notice that we can do this because r and ρ are exogenous. This assumption eliminates any inessential dynamics (including endogenous growth) and ensures a well-behaved BGP.[3] It follows then that consumption is constant:

$$c_t = c^*. \tag{4.18}$$

4.2.3 | Solving for the stock of domestic capital

FOC (4.13) says that the marginal product of (per-capita) capital is constant and equal to the interest rate on bonds. Intuitively, the marginal return on both assets is equalised. This means that capital is always at its steady state level k^*, which is defined by

$$f'(k^*) = r. \tag{4.19}$$

This means, in turn, that domestic per capita income is constant, and given by

$$y^* = f(k^*). \tag{4.20}$$

Note that the capital stock is completely independent of savings and consumption decisions, which is a crucial result of the small open economy model. One should invest in capital according to its rate of return (which is benchmarked by the return on bonds), and raise the necessary resources not out of savings, but out of debt.

4.2.4 | The steady state consumption and current account

Now that you have the level of income you should be able to compute the level of consumption. How do we do that? By solving the differential equation that is the budget constraint (4.9), which we can rewrite as

$$\dot{a}_t - ra_t = f(k^*) - rk^* - c^*, \tag{4.21}$$

using the solutions for optimal consumption and capital stock. Using our strategy of integrating factors, we can multiply both sides by e^{-rt}, and integrate the resulting equation between 0 and t:

$$a_t e^{-rt} - a_0 = \frac{c^* + rk^* - f(k^*)}{r}(e^{-rt} - 1). \tag{4.22}$$

Now evaluate this equation as $t \to \infty$. Considering the NPC and the TVC, it follows that:

$$c^* = ra_0 + f(k^*) - rk^*. \tag{4.23}$$

We can also find the optimal level of debt at each time period. It is easy to see that a_t is kept constant at a_0, from which it follows that $b_t = b_0 + k_0 - k^*$. The current account is zero. In other words, the NGM delivers a growth model with no growth, as we saw in the last chapter, and a model of the current account dynamics without current account surpluses or deficits.

Not so fast, though! We saw that the NGM did have predictions for growth outside of the BGP. Let's look at the transitional dynamics here as well, and see what we can learn.

4.2.5 | The inexistence of transitional dynamics

There are no transitional dynamics in this model: output per capita converges instantaneously to that of the rest of the world!

Suppose that initial conditions are $k_0 < k^*$ and $b_0 > 0$. But, condition (4.19) says that capital must always be equal to k^*. Hence, in the first instant, capital must jump up from k_0 to k^*. How is this accomplished? Domestic residents purchase the necessary quantity of capital (the single good) abroad and instantaneously install it. Put differently, the speed of adjustment is infinite.

How do the domestic residents pay for this new capital? By drawing down their holdings of the bond. If $\Delta k_0 = k^* - k_0$, then $\Delta b_0 = -\Delta k_0 = -\left(k^* - k_0\right)$. Note that this transaction does not affect initial net national assets, since

$$\Delta a_0 = \Delta k_0 + \Delta b_0 = \Delta k_0 - \Delta k_0 = 0. \tag{4.24}$$

An example

Suppose now that the production function is given by

$$f\left(k_t\right) = Ak_t^{\alpha}, A > 0, 0 \leq \alpha \leq 1. \tag{4.25}$$

This means that condition (4.19) is

$$\alpha A\left(k^*\right)^{\alpha-1} = r \tag{4.26}$$

so that the level of capital on the BGP is

$$k^* = \left(\frac{\alpha A}{r}\right)^{\frac{1}{1-\alpha}}, \tag{4.27}$$

which is increasing in A and decreasing in r.

Using this solution for the capital stock we can write y^* as

$$y^* = Ak^{*\alpha} = A\left(\frac{\alpha A}{r}\right)^{\frac{\alpha}{1-\alpha}} = A^{\frac{1}{1-\alpha}}\left(\frac{\alpha}{r}\right)^{\frac{\alpha}{1-\alpha}} \equiv z(A), \tag{4.28}$$

with $z(A)$ increasing in A.

It follows that consumption can be written as

$$c^* = ra_0 - rk^* + z(A) = ra_0 + (1-\alpha)z(A), \tag{4.29}$$

with $z'(A) > 0$.

4.2.6 | Productivity shocks and the current account

Suppose the economy initially has total factor productivity A^H, with corresponding optimal stock of capital $(k^*)^H$ and consumption level $(c^*)^H$. At time 0 there is an unanticipated and permanent fall in productivity from A^H to A^L, where $A^L < A^H$ (maybe because this economy produced oil, guano, or diamonds and its price has come down). This means, from (4.28), that $z(A)$ falls from $z\left(A^H\right)$ to $z\left(A^L\right)$. Capital holdings are reduced: residents sell capital in exchange for bonds, so after the shock they have $(k^*)^L < (k^*)^H$, where $(k^*)^H$ was the optimal stock of capital before the shock. Assets a_0 are unchanged on impact.

From (4.29) it follows that consumption adjusts instantaneously to its new (and lower) value:

$$(c^*)^L = ra_0 - (1-\alpha)z\left(A^L\right) < ra_0 - (1-\alpha)z\left(A^H\right) = (c^*)^H, \text{ for all } t \geq 0. \tag{4.30}$$

What happens to the current account? After the instantaneous shock, assets remain unchanged, and $\dot{b}_t$ is zero. The economy immediately converges to the new BGP, where the current account is in balance.

At this point, you must be really disappointed: don't we ever get any interesting current account dynamics from this model? Actually, we do! Consider a *transitory* fall in productivity at time 0, from A^H to A^L, with productivity eventually returning to A^H after some time T. Well, it should be clear that consumption will fall, but not as much as in the permanent case. You want to smooth consumption, and you understand that things will get back to normal in the future, so you don't have to bring it down so much now. At the same time, the capital stock does adjust down fully, otherwise its return would be below what the domestic household could get from bonds. If current output falls just as in the permanent case, but consumption falls by less, where is the difference? A simple inspection of (4.9) reveals that $\dot{b}$ has to fall below zero: it's a current-account deficit! Quite simply, residents can smooth consumption, in spite of the negative shock, by borrowing resources from abroad. Once the shock reverts, the current account returns to zero, while consumption remains unchanged. In the new BGP, consumption will remain lower relative to its initial level, and the difference will pay for the interest incurred on the debt accumulated over the duration of the shock – or more generally, the reduction in net foreign asset income.

This example underscores the role of the current account as a mechanism through which an economy can adjust to shocks. It also highlights one feature that we will see over and over again: the optimal response and resulting dynamics can be very different depending on whether a shock is permanent or transitory.

4.2.7 | Sovereign wealth funds

This stylised model actually allows us to think of other simple policy responses. Imagine a country that has a finite stock of resources, like copper.[4] Furthermore let's imagine that this stock of copper is being extracted in a way that it will disappear in a finite amount of time. The optimal program is to consume the net present value of the copper over the infinite future. So, as the stock of copper declines the economy should use those resources to accumulate other assets. This is the fiscal surplus rule implemented by Chile to compensate for the depletion of their resources. In fact, Chile also has a rule to identify transitory from permanent shocks, with the implication that all transitory increases (decreases) in the price level have to be saved (spent).

Does this provide a rationale for some other sovereign wealth funds? The discussion above suggests that a country should consume:

$$r\int_{-\infty}^{\infty} R_t e^{-rt} dt, \tag{4.31}$$

where R is the value of the resources extracted in period t. This equation says that a country should value its intertemporal resources (which are the equivalent of the a_0 above, an initial stock of assets), and consume the real return on it.

Is that how actual sovereign funds work? Well, the Norwegian sovereign fund rule, for instance, does not do this. Their rule is to spend at time t the real return of the assets accumulated until then:

$$r\int_{-\infty}^{t} R_t e^{-r(s-t)} ds. \tag{4.32}$$

This rule can only be rationalised if you expect no further discoveries and treat each new discovery as a surprise. Alternatively, one could assume that the future is very uncertain, so one does not want to commit debt ahead of time. (We will come back to this precautionary savings idea in our study of consumption in Chapter 11.) In any event, the key lesson is that studying our stylised models can help clarify the logic of existing policies, and where and why they depart from our basic assumptions.

4.3 | What have we learned?

The NGM provides the starting point for a lot of dynamic macroeconomic analysis, which is why it is one of the workhorse models of modern macroeconomics. In this chapter, we have seen how it provides us, in the context of a small open economy, with a theory of the current account. When an economy has access to international capital markets, it can use them to adjust to shocks, while smoothing consumption and maintaining the optimal capital stock. Put simply, by borrowing and lending (as reflected by the current account), the domestic economy need not rely on its own savings in order to make adjustments.

This brings to the forefront a couple of important messages. First, current account deficits (contrary to much popular perception) are not inherently bad. They simply mean that the economy is making use of resources in excess of its existing production capacity. That can make a lot of sense, if the economy has accumulated assets or otherwise expects to be more productive in the future.

Second, access to capital markets can be a very positive force. It allows economies to adjust to shocks, thereby reducing volatility in consumption. It is important to note that this conclusion is coming from a model without risk or uncertainty, without frictions in capital markets, and where decisions are being taken optimally by a benevolent central planner. We will lift some of those assumptions later in the book, but, while we will not spend much more time talking about open economies, it is important to keep in mind those caveats here as well.

Third, we have seen how the adjustment to permanent versus transitory shocks can be very different. We will return to this theme over and over again over the course of this book.

Last but not least, we have illustrated how our stylised models can nevertheless illuminate actual policy discussions. This will, again, be a recurrent theme in this book.

4.4 | What next?

The analysis of the current account has a long pedigree in economics. As the counterpart of current accounts are either capital flows or changes in Central Bank reserves it has been the subject of much controversy. Should capital accounts be liberalised? Is there a sequence of liberalisation? Can frictions in capital markets or incentive distortions make these markets not operate as smoothly and beneficially as we have portrayed here? The literature on moral hazard, the policy discussion on bailouts, and, as a result, all the discussion on sovereign debt, which is one key mechanism countries, smooth consumption over time. The presentation here follows Blanchard and Fischer (1989), but if you want to start easy you can check the textbook by Caves et al. (2007), which covers all the policy issues. Obstfeld and Rogoff (1996) is the canonical textbook in international finance. More recently, you can dwell in these discussions by checking out Vegh (2013) and Uribe and Schmitt-Grohé (2017). Last, but not least, the celebrated paper by Aguiar and Gopinath (2007) distinguishes between shocks to output and shocks to trends in output growth, showing that the latter are relevant empirically and help understand the current account dynamics in emerging economies.

Notes

[1] We should add secondary income, but we will disregard for the analysis.

[2] The fact that current accounts seem to be typically quite small relative to the size of the economy, so that savings is roughly similar to investment, is called the Feldstein-Horioka puzzle.

[3] Think about what happens, for instance, if $r > \rho$. We would have consumption increasing at a constant rate. This patient economy, with relatively low ρ, would start accumulating assets indefinitely. But in this case, should we expect that the assumption that it is a small economy would keep being appropriate? What if $r < \rho$? This impatient economy would borrow heavily to enjoy a high level of

consumption early on, and consumption would asymptotically approach zero as all income would be devoted to debt payments – not a very interesting case.

[4] Is this example mere coincidence, or related to the fact that one of us is from Chile, which is a major exporter of copper? We will let you guess.

References

Aguiar, M. & Gopinath, G. (2007). Emerging market business cycles: The cycle is the trend. *Journal of Political Economy*, *115*(1), 69–102.

Blanchard, O. & Fischer, S. (1989). *Lectures on macroeconomics*. MIT Press.

Caves, R. E., Frankel, J., & Jones, R. (2007). *World trade and payments: An introduction*. Pearson.

Obstfeld, M. & Rogoff, K. (1996). *Foundations of international macroeconomics*. MIT Press.

Uribe, M. & Schmitt-Grohé, S. (2017). *Open economy macroeconomics*. Princeton University Press.

Vegh, C. (2013). *Open economy macroeconomics in developing countries*. MIT Press.

CHAPTER 5

Endogenous growth models I: Escaping diminishing returns

We are still searching for the Holy Grail of endogenous growth. How can we generate growth within the model, and not as an exogenous assumption, as in the Neoclassical Growth Model with exogenous technological progress? Without that, we are left unable to really say much about policies that could affect long-run growth. We have mentioned two possible approaches to try and do this. First, we can assume different properties for the production function. Perhaps, in reality, there are features that allow economies to escape the limitations imposed by diminishing returns to accumulation. Second, we can endogenise the process of technological change so we can understand its economic incentives. The former is the subject of this chapter, and we will discuss the latter in the next one.

5.1 | The curse of diminishing returns

You will recall that a crucial lesson from the Neoclassical Growth Model was that capital accumulation, in and of itself, cannot sustain long-run growth in per capita income. This is because of *diminishing returns* to the use of capital, which is a feature of the neoclassical production function. In fact, not only are there diminishing returns to capital (i.e. $\frac{\partial^2 F}{\partial K^2} < 0$) but these diminishing returns are strong enough that we have the Inada condition that $\lim_{K \to \infty} \frac{\partial F}{\partial K} = 0$. Because of this, as you accumulate capital, the incentive to save and invest further will become smaller, and the amount of capital per worker will eventually cease to grow. The crucial question is: Are there any features of real-world technologies that would make us think that we can get away from diminishing returns?

5.2 | Introducing human capital

We show, in the context of the Solow model, how expanding the concept of capital accumulation can generate endogenous growth. This, however, depends on total returns to accumulation being non-diminishing.

One possibility is that the returns to accumulation are greater than we might think at first. This is because there is more to accumulation than machines and plants and bridges. For instance, we can also invest in and accumulate human capital!

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 5. 'Endogenous growth models I: Escaping diminishing returns', pp. 51–68. London: LSE Press. DOI: https://doi.org/10.31389/lsepress.ame.e

Could that allow us to escape the curse, and achieve sustainable growth? Here is where a formal model is once again required. We will do that in the simplest possible context, that of the Solow model, but now introducing a new assumption on the production function: the presence of human capital as an additional factor. To fix ideas, let's go to the Cobb-Douglas case:

$$Y = K^{\alpha} H^{\beta} (AL)^{\gamma} . \tag{5.1}$$

Note that we are assuming the technological parameter A to be of the labour-augmenting kind. It enters into the production function by making labour more effective.[1] Dividing through by L we obtain

$$\frac{Y}{L} = A^{\gamma} \left(\frac{K}{L}\right)^{\alpha} \left(\frac{H}{L}\right)^{\beta} L^{(\alpha+\beta+\gamma)-1}, \tag{5.2}$$

where $\alpha + \beta + \gamma$ is the scale economies parameter. If $\alpha + \beta + \gamma = 1$, we have constant returns to scale (CRS). If $\alpha + \beta + \gamma > 1$, we have increasing returns to scale; doubling all inputs more than doubles output.

Assume CRS for starters. We can then write the production function as

$$y = A^{\gamma} k^{\alpha} h^{\beta}, \tag{5.3}$$

where, as before, small-case letters denote per-capita variables.

5.2.1 | Laws of motion

Let us start way back in the Solow world. As in the simple Solow model, assume constant propensities to save out of current income for physical and human capital, $s_k, s_h \in (0, 1)$. Let δ be the common depreciation rate. We then have

$$\dot{K} = s_k Y - \delta K, \tag{5.4}$$

$$\dot{H} = s_h Y - \delta H, \tag{5.5}$$

and, therefore,

$$\frac{\dot{K}}{L} = s_k y - \delta k, \tag{5.6}$$

$$\frac{\dot{H}}{L} = s_h y - \delta h. \tag{5.7}$$

Recall next that

$$\frac{\dot{K}}{L} = \dot{k} + nk, \tag{5.8}$$

$$\frac{\dot{H}}{L} = \dot{h} + nh. \tag{5.9}$$

Using these expressions we have

$$\dot{k} = s_k A^{\gamma} k^{\alpha} h^{\beta} - (\delta + n) k, \tag{5.10}$$

$$\dot{h} = s_h A^{\gamma} k^{\alpha} h^{\beta} - (\delta + n) h, \tag{5.11}$$

which yield:

$$\gamma_k = \frac{\dot{k}}{k} = s_k A^\gamma k^{\alpha-1} h^\beta - (\delta + n), \tag{5.12}$$

$$\gamma_h = \frac{\dot{h}}{h} = s_h A^\gamma k^{\alpha} h^{\beta-1} - (\delta + n). \tag{5.13}$$

5.2.2 | Balanced growth path

You will recall that a BGP is a situation where all variables grow at a constant rate. From (5.12) and (5.13) (and in the absence of technological progress), we see that constant γ_k and γ_h require, respectively[2],

$$(\alpha - 1)\gamma_k + \beta\gamma_h = 0, \tag{5.14}$$

$$\alpha\gamma_k + (\beta - 1)\gamma_h = 0. \tag{5.15}$$

Substituting the second equation into the first equation yields

$$\frac{1-\alpha-\beta}{1-\beta}\gamma_k = 0. \tag{5.16}$$

But given CRS, we have assumed that $\alpha + \beta < 1$, so we must have $\gamma_k = \gamma_h = 0$. In other words, just as before, without technical progress (A constant), this model features constant per-capita capital k and constant per-capita human capital h. No growth again! Of course, we can obtain long-run growth again by assuming exogenous (labour-augmenting) technological progress, $\frac{\dot{A}}{A} = g$. Consider a BGP in which $\frac{\dot{k}}{k}$ and $\frac{\dot{h}}{h}$ are constant over time. From (5.12) and (5.13), this requires that $\frac{k}{y}$ and $\frac{h}{y}$ be constant over time. Consequently, if a BGP exists, y, k, and h, must all be increasing at the same rate. When the production function exhibits CRS, this BGP can be achieved by setting $\frac{\dot{y}}{y} = \frac{\dot{k}}{k} = \frac{\dot{h}}{h} = g$.[3] The long-run growth rate is thus independent of s_k, s_h, n or anything that policy affects, unless g is endogenised somehow. (But again, long-run levels of income do depend on these behavioural parameters.)

5.2.3 | Still looking for endogenous growth

Why is the long-run growth rate still pinned down by the exogenous rate of technological growth as in the Solow Model? CRS implies that the marginal products of K and H decline as these factors accumulate, tending to bring growth rates down. Moreover, Cobb-Douglas production functions satisfy the Inada conditions so that, in the limit, these marginal products asymptotically go to 0. In other words, CRS still keeps us in the domain of diminishing returns to capital accumulation, regardless of the fact that we have introduced human capital!

How can we change the model to make long-run growth rates endogenous (i.e., potentially responsive to policy)? You should see immediately from (5.16) that there is a possibility for a BGP, with γ_k and γ_h different from zero: if $\alpha + \beta = 1$. That is to say, if we have constant returns to capital and human capital, the reproducible factors, taken together.

It is easy to see, from (5.12) and (5.13), that in a BGP we must have

$$\frac{\dot{k}}{k} = \frac{\dot{h}}{h} \longrightarrow \frac{k^*}{h^*} = \frac{s_k}{s_h}. \tag{5.17}$$

In other words, in a BGP k and h must grow at the same rate. This is possible since diminishing returns does not set in to either factor ($\alpha + \beta = 1$). What rate of growth is this? Using (5.17) in (5.12) and (5.13) we obtain (normalizing $A = 1$ for simplicity)

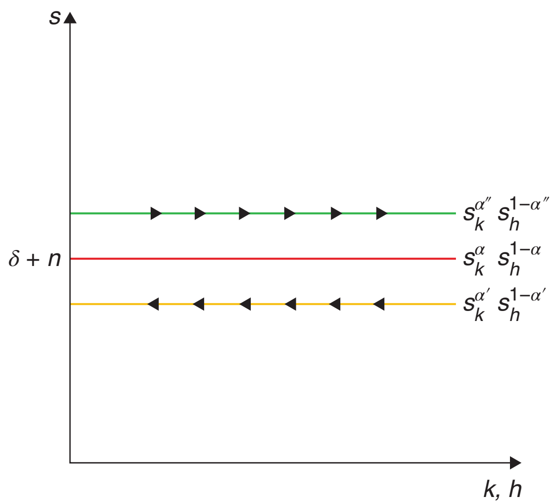
$$\frac{\dot{k}}{k} = \frac{\dot{h}}{h} = s_h \left(\frac{s_k}{s_h}\right)^{\alpha} - (\delta + n) = s_k^{\alpha} s_h^{1-\alpha} - (\delta + n). \tag{5.18}$$

The long-run (BGP) growth rate of output is

$$\frac{\dot{y}}{y} = \alpha \frac{\dot{k}}{k} + (1 - \alpha) \frac{\dot{h}}{h} = s_k^{\alpha} s_h^{1-\alpha} - (\delta + n). \tag{5.19}$$

Now s_k, s_h do affect long-run growth. If policy affects these, then policy affects growth. For instance, increasing the savings rates leads to higher growth in the long run. In other words, when we have human capital *and constant returns to reproducible factors of production*, it is possible to explain long-run growth (see Figure 5.1).

Figure 5.1 Endogenous growth



A couple of observations are in order. First, with permanent differences in growth rates across countries, the cross-national variation of per-capita incomes will blow up over time. In other words, there is no convergence in such a model. Also, if there is technical progress, growth rates will be higher.

5.3 | The AK model

We embed the notion of non-diminishing returns to accumulation into the setting of the Ramsey problem: $f(k) = Ak$. The resulting Euler equation, $\frac{\dot{c}_t}{c_t} = \sigma(A - \rho)$, displays endogenous growth. This is a very different world from the NGM: there are no transitional dynamics, policies affect long-run growth, there is no convergence, and temporary shocks have permanent effects.

The model in the previous section, just like the Solow model, was not micro-founded in terms of individual decisions. Let us now consider whether its lessons still hold in a framework with optimising individuals.

We have seen that the key aspect to obtaining long-run growth in the previous model is to have constant returns to reproducible factors when taken together. Including human capital as one such factor is but one way of generating that. To keep things as general as possible, though, we can think of all reproducible factors as capital, and we can subsume all of these models into the so-called AK model.

Consider once again a model with one representative household living in a closed economy, members of which consume and produce. There is one good, and no government. Population growth is 0, and the population is normalised to 1. All quantities (in small-case letters) are per-capita. Each consumer in the representative household lives forever.

The utility function is

$$\int_0^{\infty} \left(\frac{\sigma}{\sigma - 1}\right) c_t^{\frac{\sigma-1}{\sigma}} e^{-\rho t} dt, \ \ \rho > 0, \tag{5.20}$$

where c_t denotes consumption, ρ is the rate of time preference and σ is the elasticity of intertemporal substitution in consumption.

We have the linear production function from which the model derives its nickname:

$$Y_t = Ak_t, A > 0. \tag{5.21}$$

Again, think of household production: the household owns the capital and uses it to produce output.

The resource constraint of the economy is

$$\dot{k}_t = Ak_t - c_t. \tag{5.22}$$

5.3.1 | Solution to household's problem

The household's problem is to maximise (5.20) subject to (5.22) for given k_0. The Hamiltonian for the problem can be written as

$$H = \left(\frac{\sigma}{\sigma - 1}\right) c_t^{\frac{\sigma-1}{\sigma}} + \lambda_t \left(Ak_t - c_t\right). \tag{5.23}$$

Note c is the control variable (jumpy), k is the state variable (sticky), and λ is the costate.

First order conditions are

$$\frac{\partial H}{\partial c_t} = c_t^{-\frac{1}{\sigma}} - \lambda_t = 0, \tag{5.24}$$

$$\dot{\lambda}_t = -\frac{\partial H}{\partial k_t} + \rho\lambda_t = -A\lambda_t + \rho\lambda_t, \tag{5.25}$$

$$\lim_{t\to\infty}\left(k_t\lambda_t e^{-\rho t}\right) = 0. \tag{5.26}$$

This last expression is, again, the transversality condition (TVC).

5.3.2 | At long last, a balanced growth path with growth

Take (5.24) and differentiate both sides with respect to time, and divide the result by (5.24) to obtain

$$-\frac{1}{\sigma}\frac{\dot{c}_t}{c_t} = \frac{\dot{\lambda}_t}{\lambda_t}. \tag{5.27}$$

Multiplying through by $-\sigma$, (5.27) becomes

$$\frac{\dot{c}_t}{c_t} = -\sigma\left(\frac{\dot{\lambda}_t}{\lambda_t}\right). \tag{5.28}$$

Finally, using (5.25) in (5.28) we obtain

$$\frac{\dot{c}_t}{c_t} = \sigma\left(A - \rho\right), \tag{5.29}$$

which is the Euler equation. Note that here we have $f'(k) = A$, so this result is actually the same as in the standard Ramsey model. The difference is in the nature of the technology, as now we have constant returns to capital.

Define a BGP once again as one in which all variables grow at a constant speed. From (5.22) we get

$$\frac{\dot{k}_t}{k_t} = A - \frac{c_t}{k_t}. \tag{5.30}$$

This implies that capital and consumption must grow at the same rate – otherwise we wouldn't have $\frac{\dot{k}_t}{k_t}$ constant. And since $y_t = Ak_t$, output grows at the same rate as well. From (5.29) we know that this rate is $\sigma\left(A - \rho\right)$. Hence,

$$\frac{\dot{c}_t}{c_t} = \frac{\dot{k}_t}{k_t} = \frac{\dot{y}_t}{y_t} = \sigma\left(A - \rho\right). \tag{5.31}$$

Note, there will be positive growth only if $A > \rho$ that is, only if capital is sufficiently productive so that it is desirable to accumulate it.

Second, from (5.30) we see that along a BGP we must have

$$y_t - c_t = \sigma(A-\rho)k_t \Rightarrow c_t = \left[(1-\sigma)A + \sigma\rho\right]k_t = \left[\frac{(1-\sigma)A + \sigma\rho}{A}\right]y_t. \tag{5.32}$$

In words, consumption is proportional to capital. Or, put differently, the agent consumes a fixed share of output every period. Notice that this is much like the assumption made in Solow. If s is the savings rate, here $1 - s = \frac{(1-\sigma)A+\sigma\rho}{A}$, or $s = \sigma\left(\frac{A-\rho}{A}\right)$. The difference is that this is now optimal, not arbitrary.

There are no transitional dynamics: the economy is always on the BGP.

5.3.3 | Closing the model: The TVC and the consumption function

We must now ask the following question. Are we sure the BGP is optimal? If $A > \rho$, the BGP implies that the capital stock will be growing forever. How can this be optimal? Would it not be better to deplete the capital stock? More technically, is the BGP compatible with the TVC? Since we did not use it in constructing the BGP, we cannot be sure. So, next we check the BGP is indeed optimal in that, under some conditions, it does satisfy the TVC.

Using (5.24) the TVC can be written as

$$\lim_{t\to\infty}\left(k_t c_t^{-\frac{1}{\sigma}} e^{-\rho t}\right) = 0. \tag{5.33}$$

Note next that equation (5.29) is a differential equation which has the solution

$$c_t = c_0 e^{\sigma(A-\rho)t}. \tag{5.34}$$

Combining the last two equations the TVC becomes

$$\lim_{t\to\infty}\left(k_t c_0^{-\frac{1}{\sigma}} e^{-At}\right) = 0. \tag{5.35}$$

From the solution to expression (5.31) we have

$$k_t = k_0 e^{\sigma(A-\rho)t}. \tag{5.36}$$

Using this to eliminate k_T, the TVC becomes

$$\lim_{t\to\infty}\left(k_0 c_0^{-\frac{1}{\sigma}} e^{\sigma(A-\rho)t} e^{-At}\right) = \lim_{t\to\infty}\left(k_0 c_0^{-\frac{1}{\sigma}} e^{-\theta t}\right) = 0, \tag{5.37}$$

where

$$\theta \equiv (1-\sigma)A + \sigma\rho. \tag{5.38}$$

Hence, for the TVC we need $\theta > 0$, which we henceforth assume. Note that with logarithmic utility ($\sigma = 1$), $\theta = \rho$.

5.3.4 | The permanent effect of transitory shocks

In the AK model, as we have seen, growth rates of all pertinent variables are given by $\sigma(A-\rho)$. So, if policy can affect preferences (σ, ρ) or technology (A), it can affect growth.

If it can do that, it can also affect levels. From the production function, in addition to (5.31) and (5.32), we have

$$k_t = k_0 e^{\sigma(A-\rho)t}, \tag{5.39}$$

$$y_t = A k_0 e^{\sigma(A-\rho)t}, \tag{5.40}$$

$$c_t = [(1-\sigma)A + \sigma\rho]\, k_0 e^{\sigma(A-\rho)t}. \tag{5.41}$$

Clearly, changes in σ, ρ and A matter for the levels of variables.

Notice here that there is no convergence in per capita incomes whatsoever. Countries with the same σ, ρ, and A retain their income gaps forever.[4]

Consider the effects of a sudden increase in the marginal product of capital A, which suddenly and unexpectedly rises (at time $t = 0$), from A to $A' > A$. Then, by (5.31), the growth rate of all variables immediately rises to $\sigma \left(A' - \rho\right)$.

What happens to the levels of the variables? The capital stock cannot jump at time 0, but consumption can. The instant after the shock ($t = 0^+$), it is given by

$$c_{0^+} = \left[(1-\sigma) A' + \sigma\rho\right] k_{0^+} > c_0 = [(1-\sigma) A + \sigma\rho]\, k_{0^+}, \tag{5.42}$$

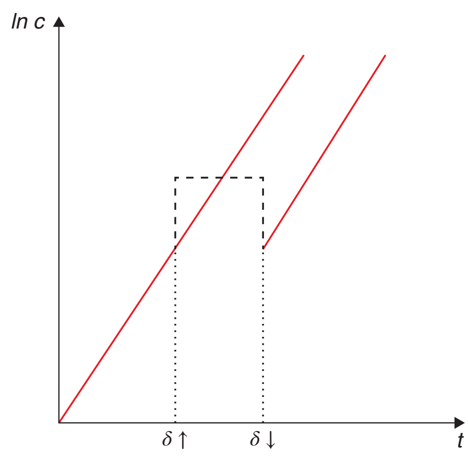
where $k_{0^+} = k_0$ by virtue of the sticky nature of capital.

So, consumption rises by $(1-\sigma)\left(A' - A\right) k_0$. But, output rises by $\left(A' - A\right) k_0$. Since output rises more than consumption, growth picks up right away.

It turns out that the AK model has very different implications from the Neoclassical Growth Model when it comes to the effects of transitory shocks. To see that, consider a transitory increase in the discount factor, i.e. suppose ρ increases for a fixed interval of time; for simplicity, assume that the new ρ is equal to A.

Figure 5.2 shows the evolution of the economy: the transitory increase in the discount rate jolts consumption, bringing growth down to zero while the discount factor remains high. When the discount factor reverts, consumption decreases, and growth restarts. But there is a permanent fall in the level of output relative to the original path. In other words, there is full persistence of shocks, even if the shock itself is temporary. You may want to compare this with the Neoclassical Growth Model trajectories (Figure 5.3), where there is catch-up to the original path and there are no long-run effects.

Figure 5.2 Transitory increase in discount rate

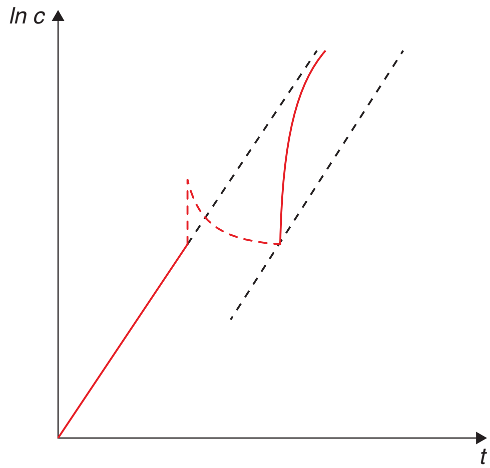


5.3.5 | In sum

In AK models of endogenous growth:

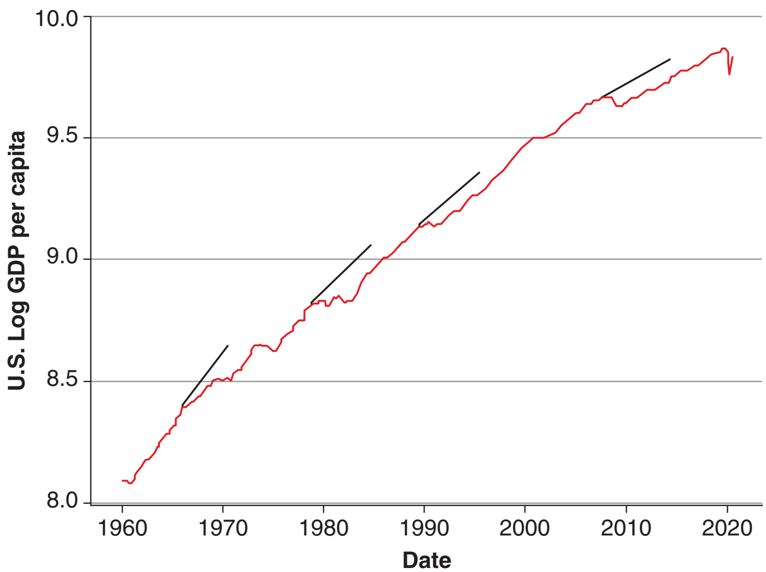
1. There is no transitional dynamics;
2. Policies that affect the marginal product of capital (e.g. taxes) do affect growth;
3. There is no convergence;
4. Even temporary policies have permanent effects.

Figure 5.3 Comparison with Solow model



These results are surprising and were initially presented by Romer (1987), as part of the contributions that eventually won him the Nobel Prize in Economics in 2018. You have to admit that this is very different from the world of diminishing returns depicted by the NGM. Now look at the graph from the U.S. recovery after the Great Recession of 2008/2009 and notice the similarities with the dynamics of the AK model: no return to the previous growth trend. The graph even suggests that it is not the first time a pattern like this plays out. Maybe this model is onto something, after all.

Figure 5.4 U.S. real GDP and extrapolated trends



5.4 | Knowledge as a factor of production

We argue that knowledge explains why accumulation may not face diminishing returns. We develop different models of how this may happen (learning-by-doing, specialisation). In the process, we show that in a world with non-diminishing returns to accumulation (and hence increasing returns to scale), the decentralised equilibrium need not be efficient: growth will be lower than than the social optimum as private incentives to accumulate knowledge are below the social returns.

We have seen that the key to obtaining long-run growth in our models is to get constant returns in the reproducible factors. But this begs the question: why do we think that this would actually be the case in reality?

As we have seen, a world of constant returns to reproducible factors is, actually, a world with *increasing returns to scale* (IRS) – after all, there is at least labour and technology in the production functions as well. But, this is a problem because IRS implies that our simple model with perfect competition doesn't really work anymore.

To see why, note that with perfect competition, each factor of production gets paid its marginal product – you know that from Econ 101. However, if the production function is

$$F(A, X), \tag{5.43}$$

where X has constant returns, then we have

$$F(A, X) < A\frac{\partial F}{\partial A} + X\frac{\partial F}{\partial X}. \tag{5.44}$$

There is not enough output to pay each factor their marginal productivity!

We had sidestepped this discussion up to this point, assuming that technology was there and was left unpaid. But now the time has come to deal with this issue head-on.

In doing so, we will build a bridge between what we have learned about accumulation and what we have talked about when referring to productivity. The crucial insight again is associated with Paul Romer, and can be summarised in one short sentence: economies can grow by accumulating "knowledge".

But what drives the accumulation of knowledge? Knowledge is a tricky thing because it is difficult to appropriate, i.e. it has many of the properties of a public good. As you may remember, the two distinguishing characteristics of any good are

- Rivalry ⟶ if I use it you can't.
- Excludability ⟶ I can prevent you from using it.

Private goods are rival and excludable, pure public goods are neither. Technology/knowledge is peculiar because it is non-rival, although excludable to some extent (with a patent, for example).

The non-rivalry of knowledge immediately gives rise to increasing returns. If you think about it, knowledge is a fixed cost: in order to produce one flying car, I need one blueprint for a flying car, but I don't need a second blueprint to build a second unit of that flying car. In other words, one doesn't need to double all inputs in order to double output.

This complicates our picture. If factors of production cannot be paid their marginal returns, and there is not enough output to pay them all, then how is the accumulation of knowledge paid for? Here are the options:

1. A is public and provided by the government;
2. Learning by doing (i.e. externalities, again);
3. Competitive behaviour is not preserved.

We will not deal much with #1 (though it is clear that in the areas where research has more externalities and is less excludable, as in basic research, there is a larger participation of the public sector), but we will address some relevant issues related to #2 (here) and #3 (next chapter).

5.4.1 | Learning by doing

This was first suggested by Romer (1987). The idea is that you become better at making stuff as you make it: knowledge is a by-product of production itself. This means that production generates an externality. If each firm does not internalise the returns to the knowledge they generate and that can be used by others, firms still face convex technologies even though there are increasing returns at the level of the economy. It follows that competitive behaviour can be preserved.

Let us model this with the following production function,

$$y = Ak^{\alpha}\bar{k}^{\eta}, \tag{5.45}$$

where $\bar{k}$ is the stock of knowledge (past investment). Given this we compute the (private) marginal product of capital and the growth rate:

$$f'(k) = A\alpha k^{\alpha-1}\bar{k}^{\eta} = A\alpha k^{\alpha+\eta-1}, \tag{5.46}$$

$$\gamma_c = \sigma\left(A\alpha k^{\alpha+\eta-1} - \rho\right). \tag{5.47}$$

We have endogenous growth if $\alpha + \eta \geq 1$. Notice that we need CRS in the reproducible factors, and, hence, sufficiently strong IRS. It is not enough to have IRS; we need that $\eta \geq 1 - \alpha$.

For a central planner who sees through the learning-by-doing exercise:

$$f(k) = Ak^{\alpha+\eta}, \tag{5.48}$$

$$f'(k) = (\alpha + \eta)\, Ak^{\alpha+\eta-1}, \tag{5.49}$$

$$\gamma_p > \gamma_c. \tag{5.50}$$

It follows that the economy does not deliver the right amount of growth. Why? Because of the externality: private agents do not capture the full social benefit of their investment since part of it spills over to everyone else. This is a crucial lesson of the endogenous growth literature. Once we introduce IRS, there will typically be a wedge between the decentralised equilibrium and the optimal growth rate.

5.4.2 | Adam Smith's benefits to specialisation

The second story, (from Romer 1990), suggests that economies can escape diminishing returns by increasing the range of things they produce, an idea akin to Adam Smith's suggestion that specialisation increases productivity. Suppose the production function could use a continuum of potential inputs.

$$Y(X, L) = L^{1-\alpha} \int_0^\infty X(i)^\alpha \, di. \tag{5.51}$$

But not all varieties are produced. Let's say only the fraction $[0, M]$ are currently available. Say the average cost of production of each intermediate unit is 1, this implies that of each unit I will use

$$X(i) = \bar{X} = \frac{Z}{M}, \tag{5.52}$$

where Z are total resources devoted to intermediate inputs. So, this yields

$$Y = L^{1-\alpha} M \left(\frac{Z}{M}\right)^\alpha = L^{1-\alpha} Z^\alpha M^{1-\alpha}. \tag{5.53}$$

Note that we can write $Z = M\bar{X}$, so an expansion in Z can be accomplished by increasing M, the number of varieties, or increasing $\bar{X}$, the amount of each variety that is used. In other words, you can either pour more resources into what you already do, or into doing different things. We can thus write

$$Y = L^{1-\alpha} (M\bar{X})^\alpha M^{1-\alpha} = L^{1-\alpha} \bar{X}^\alpha M. \tag{5.54}$$

Lo and behold: increasing $\bar{X}$ encounters diminishing returns ($\alpha < 1$), but that is *not* the case when one increases M. In other words, specialisation prevents diminishing returns. Choosing units appropriately, we can have

$$M = Z. \tag{5.55}$$

But this then yields

$$Y = L^{1-\alpha} Z. \tag{5.56}$$

If $\dot{Z} = Y - C$ we are done: we are back to the AK model!

A nice example of the power of diversification in the production function is obtained in Gopinath and Neiman (2014), where they use Argentina's crisis of 2001/2002, which restricted access of firms to intermediate inputs, to estimate a large impact on productivity.

We should model next how the private sector will come up with new varieties (R&D). This will typically involve non-competitive behaviour: one will only invest in R&D if there is a way of recouping that investment (e.g. patents, monopoly power). This will also lead to a wedge between the optimal growth rate and the one that is delivered by the decentralised equilibrium: monopolies will undersupply varieties. But, careful: this will not always be so. In fact, we will develop a model where monopolies will oversupply varieties as well! At any rate, we will look at this in a bit more detail in the next chapter.

In the meantime, note in particular that these wedges introduce a potential role for public policy. For instance, if there is undersupply of varieties, one could introduce a subsidy to the purchase of intermediate inputs so that producers wouldn't face monopoly prices.

5.5 | Increasing returns and poverty traps

We digress into how a specific kind of increasing returns can be associated with the existence of poverty traps: situations where economies are stuck in a stagnating equilibrium when a better one would be available with an injection of resources. We discuss whether poverty traps are an important feature in the data and policy options to overcome them.

We have just argued that the presence of IRS (associated with non-diminishing returns to accumulation) is a key to understanding long-run growth. It turns out that the presence of (certain kinds of) IRS can also explain the condition of countries that seem to be mired in poverty and stagnation – as captured by the idea of poverty traps.

The concept of a poverty trap describes a situation in which some countries are stuck with stagnant growth and/or low levels of income per capita, while other (presumably similar) countries race ahead. The key for the emergence of this pattern is the presence of IRS, at least for a range of capital-labour ratios. The idea is as old as Adam Smith, but Rosenstein-Rodan (1943), Rosenstein-Rodan (1961), Singer (1949), Nurkse (1952), Myrdal and Sitohang (1957) and Rostow (1959) appropriated it for development theory. They argued that increasing returns only set in after a nation has achieved a particular threshold level of output per capita. Poor countries, they argued, were caught in a poverty trap because they had been hitherto unable to push themselves above that threshold. The implication is that nations that do not manage to achieve increasing returns are left behind. Those that do take off into a process of growth that leads to a steady state with higher standards of living (or maybe even to never-ending growth). You should keep in mind that, while the idea of poverty traps, and the calls for "big push" interventions to lift countries above the threshold that is needed to escape them, have been around for quite a while, they are still very much in the agenda. See for instance, Sachs (2005). Of course this view has plenty of critics as well – on that you may want to check Easterly (2001)'s book, which provides a particularly merciless critique.

Let's develop one version for a story generating poverty traps based on a simple modification of the Solow model highlighting the role of increasing returns in the production function. This makes the argument in the simplest possible fashion. You can refer to the paper by Kraay and McKenzie (2014) for a discussion of what could generate this sort of increasing returns. For instance, there could be fixed costs (lumpy investments) required to access a better technology (coupled with borrowing constraints). They also tell stories based on savings behaviour, or nutritional traps, among others.

5.5.1 | Poverty trap in the Solow model

Recall that, in per capita terms, the change in the capital stock over time is given by

$$\dot{k} = s \cdot f(k) - (n + \delta) \cdot k. \tag{5.57}$$

The key to generating growth traps in the Solow model is assuming a particular shape to the production function. In particular, we assume a (twice-continuously differentiable) function such that

$$f''(k) = \begin{cases} < 0 \text{ if } 0 < k < k_a \\ > 0 \text{ if } k_a < k < k_b \\ < 0 \text{ if } \quad k > k_b. \end{cases} \tag{5.58}$$

The key is that the production function $f(k)$ has a middle portion where it exhibits increasing returns to scale.

Notice that the term $\frac{sf(k)}{k}$, crucial for the dynamics of the Solow model, has a derivative equal to

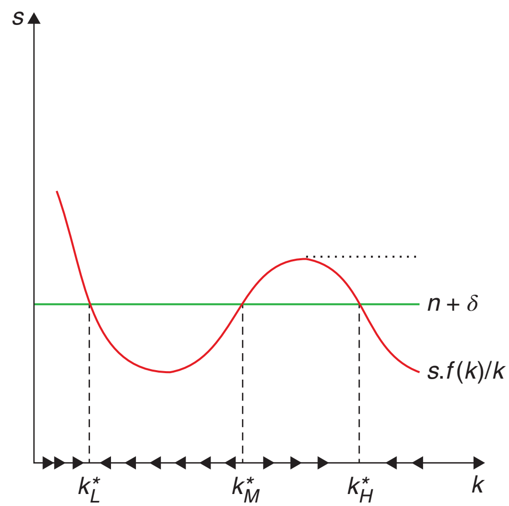
$$\frac{\partial \frac{sf(k)}{k}}{\partial k} = \frac{skf'(k) - sf(k)}{k^2} = \frac{sf'(k)}{k}\left(1 - \frac{f(k)}{kf'(k)}\right). \tag{5.59}$$

This derivative can only be zero whenever $f''(k) = 0$, which by (5.58) happens when $k = k_a$ and $k = k_b$.[5] It can also be shown that

$$\frac{\partial^2 \frac{sf(k)}{k}}{\partial k^2} = \begin{cases} > 0 \;\; if \;\; k = k_a \\ < 0 \;\; if \;\; k = k_b. \end{cases} \tag{5.60}$$

It follows that the function $\frac{sf(k)}{k}$ has the shape depicted in Figure 5.5.

Figure 5.5 Multiple equilibria in the Solow model



The dynamic features of this system, including the possibility of a poverty trap, can be read from the diagram directly. We have three steady states, at k^*_L, k^*_M and k^*_H. Of these, k^*_L and k^*_H are stable, while k^*_M is unstable. The implication is that if a country begins with a capital-labor ratio that is below k^*_M, then it will inexorably approach the steady state ratio k^*_L. If its initial capital-labour ratio is above k^*_M, then it will approach the much better steady state at k^*_H. The capital-labour ratio k^*_M, then, is the threshold capital stock (per capita) that a nation has to reach to take off and achieve the higher steady state.

Notice that in the end different countries may be at different steady state ratios, but they still exhibit identical growth rates (equal to zero). In Figure 5.5, a poor economy at steady state k^*_L and a rich economy at steady state k^*_H experience the same growth rates of aggregate variables and no growth in per capita variables. Notice, however, that the poor economy has per-capita income of $f\left(k^*_L\right)$ and the rich economy has per capita income of $f\left(k^*_H\right)$, which means that residents of the poor economy only get to enjoy consumption of magnitude $(1 - s)f\left(k^*_L\right)$, while residents of the rich economy enjoy the higher level $(1-s)\,f\left(k_H^*\right)$. Hence, differences in initial conditions imply lasting differences in consumption and welfare among economies that are fundamentally identical. (Note that the production function $f(k)$, the savings rate s, the population growth rate n, and the depreciation rate δ are the same across these economies.)

5.5.2 | Policy options to overcome poverty traps

There are a few alternative policy options for nations caught in a poverty trap. The first is to temporarily increase the savings rate. Consider Figure 5.5 and suppose that we have a country with savings rate s_1 stuck at the stagnant steady state ratio k_L^*. A rise in the savings rate will result in a situation where there is only one stable steady state ratio at a high level of k^*. Maintaining the higher savings rate for a while, the nation will enjoy a rapid rise in the capital-labour ratio towards the new steady state. However, it need not maintain this savings rate forever. Once the capital-labour ratio has gone past k_M^*, it can lower the savings rate back down to s_1. Then the country is within the orbit of the high capital-labour ratio k_H^*, and will move inexorably towards it by the standard properties of Solow adjustment. Thus, a temporary rise in the savings rate is one way for a nation to pull itself out of the poverty trap.

Similarly, another way of escaping this poverty trap is to temporarily lower the population growth rate. A nation stuck at k_L^* could move the horizontal schedule down by decreasing population growth temporarily, thereby leaving a very high k^* as the only steady-state capital-labour ratio. The old population growth can be safely restored once the Solovian dynamics naturally push the economy above k_M^*.

There is an obvious third possibility, beyond the scope of the country and into the realm of the international community, to provide a country that is mired in a poverty trap with an injection of capital, through aid, that increases its capital stock past the threshold level. This is the big push in aid advocated by some economists, as well as many politicians, multilateral organisations, and pop stars.

In all of these cases, you should note the *permanent* effects of *temporary* policy. You will recall that this is a general feature of growth models with increasing returns, and this illustrates the importance of this aspect for designing policy.

5.5.3 | Do poverty traps exist in practice?

While many people believe poverty traps are an important phenomenon in practice – thereby providing justification for existing aid efforts – the issue is very controversial. Kraay and McKenzie (2014) consider the evidence, and come down on the skeptical side.

First, they argue that the kind of income stagnation predicted by poverty trap models are unusual in the data. The vast majority of countries have experienced positive growth over recent decades, and low-income countries show no particular propensity for slower growth. Since standard models predict a threshold above which a country would break free from the trap, that indicates that most countries would have been able to do so.

Second, they argue that the evidence behind most specific mechanisms that have been posited to generate poverty traps is limited. For instance, when it comes to the fixed cost story we have mentioned, it seems that for the most part individuals don"t need a lot of capital to start a business, and the amount of capital needed to start a business appears relatively continuous.

This doesn't mean, however, as they recognise, that poverty traps cannot explain the predicament of *some* countries, regions, or individuals. Being stuck in a landlocked country in an arid region is actually terrible! Also, we shouldn't conclude from the relatively sparse evidence that aid, for instance, is bad or useless. Poverty may be due to poor fundamentals, and aid-financed investments can help improve these fundamentals. But, it should temper our view on what we should expect from these policy interventions.

5.6 | What have we learned?

We have seen that long-term growth is possible when accumulation is not subject to diminishing returns, and that this entails a world where there are increasing returns to scale. We have also argued that one key source of these increasing returns to scale, in practice, is the accumulation of knowledge: you do not have to double knowledge in order to double output. This in turn requires us to think about what drives knowledge accumulation, and we have seen a couple of alternative stories (learning-by-doing, specialisation) that help us think that through.

Very importantly, we have seen that a world of increasing returns is one that is very different from the standpoint of policy. There is no convergence – we shouldn't expect poor countries to catch up with rich countries, even when they have the same fundamental parameters. By the same token, temporary shocks have permanent consequences. This has disheartening implications, as we shouldn't expect countries to return to a pre-existing growth trend after being hit by temporary negative shocks. But it also has more cheerful ones as temporary policy interventions can have permanent results.

We have also seen how these lessons can be applied to a specific case of increasing returns, which can generate poverty traps. Whether such traps are widespread or not remains a source of debate, but the concept nevertheless illustrates the powerful policy implications of increasing returns.

5.7 | What next?

To learn more about the endogenous growth models that we have started to discuss here, the book by Jones and Vollrath (2013) provides an excellent and accessible overview. Barro and Sala-i-Martin (2003) also covers a lot of the ground at a higher technical level that should still be accessible to you if you are using this book – it's all about the dynamic optimisation techniques we have introduced here.

For a policy-oriented and non-technical discussion on growth, an excellent resource is Easterly (2001). As we have mentioned, he is particularly skeptical when it comes to big push aid-based approaches. On poverty traps, it is worth noting that there are many other stories for sources of increasing returns of the sort we discussed. A particularly interesting one is studied by Murphy et al. (1989), which formalises a long-standing argument based on demand externalities (e.g. Rosenstein-Rodan (1943)) and investigates the conditions for their validity. This is a remarkable illustration of how helpful it is to formally model arguments. Another powerful story for increasing returns (and possible traps) comes from Diamond (1982), which studies how they can come about when market participants need to search for one another, generating the possibility of coordination failures. We will return to related concepts later in the book, when discussing unemployment (Chapter 16).

Notes

[1] Again, you should be able to see quite easily that in a Cobb-Douglas production function it doesn't really matter if we write $Y = A_1 K^\alpha H^\beta (L)^\gamma$ or $Y = K^\alpha H^\beta \left(A_2 L\right)^\gamma$; it is just a matter of setting

$A_1 \equiv A_2^{\gamma}$, which we can always do. It is important for the existence of a BGP that technology be labour-augmenting – though this is a technical point that you shouldn't worry about for our purposes here. You can take a look at Barro and Sala-i-Martin (2003) for more on that.

2 Take logs and derive with respect to time.

3 Check the math! Hint: log-differentiate (5.1).

4 Here, for simplicity, we have set population growth n and depreciation δ to zero. They would also matter for levels and rates of growth of variables. In fact, introducing depreciation is exactly equivalent to reducing A – you should try and check that out!

5 Recall that the function is twice-continuously differentiable, such that f'' has to be zero at those points. To see why $f''(k) = 0$ implies that (5.59) is equal to zero, recall from Econ 101 that "marginal product is less (more) than the average product whenever the second derivative is negative (positive)". It's all tied to Euler's homogenous function theorem, which is also behind why factors cannot be paid their marginal products when there are increasing returns to scale. As usual in math, it's all in Euler (or almost).

References

Barro, R. J. & Sala-i-Martin, X. (2003). *Economic growth* (2nd ed.). MIT press.

Diamond, P. A. (1982). Aggregate demand management in search equilibrium. *Journal of Political Economy*, *90*(5), 881–894.

Easterly, W. (2001). *The elusive quest for growth: Economists' adventures and misadventures in the tropics*. MIT press.

Gopinath, G. & Neiman, B. (2014). Trade adjustment and productivity in large crises. *American Economic Review*, *104*(3), 793–831.

Jones, C. I. & Vollrath, D. (2013). *Introduction to economic growth*. WW Norton & Company, Inc. New York, *NY*.

Kraay, A. & McKenzie, D. (2014). Do poverty traps exist? Assessing the evidence. *Journal of Economic Perspectives*, *28*(3), 127–48.

Murphy, K. M., Shleifer, A., & Vishny, R. W. (1989). Industrialization and the big push. *Journal of Political Economy*, *97*(5), 1003–1026.

Myrdal, G. & Sitohang, P. (1957). *Economic theory and under-developed regions*. London: Duckworth.

Nurkse, R. (1952). Some international aspects of the problem of economic development. *The American Economic Review*, *42*(2), 571–583.

Romer, P. M. (1987). Growth based on increasing returns due to specialization. *The American Economic Review*, *77*(2), 56–62.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, *98*(5, Part 2), S71–S102.

Rosenstein-Rodan, P. N. (1943). Problems of industrialisation of eastern and south-eastern Europe. *The Economic Journal*, *53*(210/211), 202–211.

Rosenstein-Rodan, P. N. (1961). Notes on the theory of the 'big push'. *Economic development for Latin America* (pp. 57–81). Springer.

Rostow, W. W. (1959). The stages of economic growth. *The Economic History Review*, *12*(1), 1–16.

Sachs, J. (2005). *The end of poverty: How we can make it happen in our lifetime*. Penguin UK.

Singer, H. W. (1949). Economic progress in underdeveloped countries. *Social Research*, 1–11.

CHAPTER 6

Endogenous growth models II: Technological change

As we've seen, the challenge of the endogenous growth literature is how to generate growth within the model, and not simply as an assumption as in the Neoclassical Growth Model (NGM) with exogenous technological progress. The basic problem with the NGM (without exogenous technological progress) was that the incentives to capital accumulation decreased with the marginal product of capital. So, if we are to have perpetual growth we need a model that somehow gets around this issue. To do so, the literature has gone two ways. One is to change features of the production function or introduce additional factors that are complementary to the factors that are being accumulated in a way that keep the incentives to accumulation strong. The other alternative is to endogenise technological change. The first approach was the subject of the previous chapter, this chapter will focus on the second one.

Our final discussion in the previous chapter already hinted at the issues that arise when endogenising technological change. Most crucially, knowledge or ideas have many of the properties of a public good. In particular, ideas might be (or be made) *excludable* (e.g. using patents or secrecy), but they are distinctly *non-rival*. Because of that, there is a big incentive to free-ride on other people's ideas – which is a major reason why governments intervene very strongly in the support of scientific activities. We have already looked at stories based on externalities (from learning-by-doing) and specialisation. We have also seen how they give rise to a wedge between the decentralised equilibrium and the optimal rate of growth.

In this chapter, we will take this discussion further by properly studying models where technological change emerges endogenously, through firms purposefully pursuing innovation. This is not only for the pure pleasure of solving models – though that can also be true, if you are so inclined! In fact, we will be able to see how the incentives to innovate interplay with market structure. This in turn opens a window into how the links between policy domains such as market competition, intellectual property rights, or openness to trade are fundamentally related to economic growth. We will also see that it may be the case (perhaps surprisingly) that technological innovation – and, hence economic growth – may be too fast from a social welfare perspective.

There are two ways of modelling innovation: one where innovation creates additional varieties, and another where new products sweep away previous versions in a so-called quality ladder. In the product variety model, innovation introduces a new variety but it does not (fully) displace older alternatives, like introducing a new car model or a new type of breakfast cereal. This is very much along the lines of

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 6. 'Endogenous growth models II: Technological change', pp. 69–86. London: LSE Press. DOI: https://doi.org/10.31389/lsepress.ame.f

the model of specialisation that we have already seen (and to which we will return later in the chapter). In the quality ladder approach (also known as the Schumpeterian model), a new variety is simply better than old versions of the same thing and displaces it fully. (Schumpeter famously talked about firms engaging in "creative destruction" – this is what gives this approach its name.) Examples we face everyday: typewriters wiped out by word-processing computer software; the USB keys phasing out the older 5.4-inch diskette, and in turn later being displaced by cloud storage; VCRs displaced by DVD players, and eventually by online streaming; lots of different gadgets being killed by smartphones, and so on. We will develop two standard models of innovation, one for each version, and discuss some of the most important among their many implications for policy.

6.1 | Modelling innovation as product specialisation

Following up on the previous chapter, we develop a full-fledged model of innovation through the development of new product varieties. It highlights a few important points: the role of monopoly profits in spurring the pursuit of innovation by firms, and the presence of scale effects (larger economies grow faster).

We start with this version because it will be familiar from the previous chapter. It is the model of innovation through specialisation, from Romer (1990). While we then left unspecified the process through which innovation takes place – where did the new varieties come from, after all? – we will now take a direct look at that.

Let's consider a slightly different version of the production function we posited then

$$Y(X) = \left[\int_0^M X(i)^\alpha \, di\right]^{\frac{1}{\alpha}}, \tag{6.1}$$

where again $X(i)$ stands for the amount of intermediate input of variety i, and M is the range of varieties that are currently available. Recall that we treat each sector i as infinitesimally small within a continuum of sectors. We are leaving aside the role of labour in producing final output for simplicity. Instead, we will assume that labour is used in the production of intermediate inputs on a one-for-one basis so that $X(i)$ also stands for the amount of labour used to produce that amount of intermediate input.[1]

How are new varieties developed? First of all, we must devote resources to producing new varieties – think about this as the resources used in the R&D sector. To be more concrete, let's say we need workers in the R&D sector, which we will denote as Z_M, and workers to produce intermediate inputs, which we will label Z to follow the notation from the previous chapter, where we had (somewhat vaguely) used that designation for the total resources devoted to intermediate inputs. It follows from this that $Z \equiv \int_0^M X(i)di$. To pin down the equilibrium, we will posit a labour market-clearing condition: $Z_M + Z = L$, the total labour force in the economy, which we will assume constant. We will also take Z_M (and, hence, Z) to be constant.[2] We will assume that the production of new varieties is linear in R&D labour, and proportional to the existing stock of varieties, according to

$$\dot{M}_t = BZ_M M_t. \tag{6.2}$$

Note also that we can use in (6.1) the fact that each symmetric intermediate sector in equilibrium will use $X(i) = \bar{X} = \frac{Z}{M}$, given the definition of Z, just as in the previous chapter. This means we can write

$Y_t = M^{\frac{1-\alpha}{\alpha}} Z$. Given that Z is constant, it follows that Y grows at $\frac{1-\alpha}{\alpha}$ times the growth rate of M, and, hence (using (6.2)) the growth rate of Y is $\frac{1-\alpha}{\alpha} BZ_M$. It follows that, to figure out the growth rate of the economy, we need to figure out the amount of resources devoted to producing new varieties, Z_M. In short, just as in the previous chapter, economic growth comes from the amount of resources devoted to the R&D sector, which is what drives innovation.

So we need to figure out what determines Z_M. For that, we need to start by positing the market structure in this economy, in terms of the intermediate inputs, final output, and the production of varieties. On the first, we assume that each variety is produced by a monopolist that holds exclusive rights to it. The final output is then produced by competitive firms that take the price of inputs as given as well. What would one of these competitive firms do? They would try to minimise the cost of producing each unit of output at any given point in time. If $p(i)$ is the price of variety i of the intermediate input, this means choosing $X(i)$ to minimise:

$$\int_0^M p(i) X(i)\, di, \tag{6.3}$$

subject to $\left[\int_0^M X(i)^\alpha \, di\right]^{\frac{1}{\alpha}} = 1$, that is, the unit of final output. The FOC for each $X(i)$ is[3]

$$p(i) = \lambda X(i)^{\alpha-1}, \tag{6.4}$$

where λ is the corresponding Lagrange multiplier. This yields a downward-sloping demand curve for the monopolist producing intermediate input i:

$$X(i) = \left[\frac{\lambda}{p(i)}\right]^{\frac{1}{1-\alpha}}. \tag{6.5}$$

You will know from basic microeconomics – but can also easily check! – that this is a demand function with a constant elasticity equal to $\varepsilon \equiv \frac{1}{1-\alpha}$.

As for the R&D sector: there is free entry into the development of new varieties such that anyone can hire R&D workers, and take advantage of (6.2), without needing to compensate the creators of previous varieties. Free entry implies that firms will enter into the sector as long as it is possible to obtain positive profits. To determine the varieties that will emerge, we thus need to figure out what those profits are. It will take a few steps, but it will all make sense!

First, consider that if you create a new variety of the intermediate input, you get perpetual monopoly rights to its production. A profit-maximising monopolist facing a demand curve with constant elasticity ε will choose to charge a price equal to $\frac{\varepsilon}{\varepsilon-1}$ times the marginal cost, which in our case translates into the marginal cost divided by α. Since you have to use one worker to produce one unit of the intermediate input, the marginal cost is equal to the wage, and the profit per unit will be given by $\left[\frac{w_t}{\alpha} - w_t\right] = \frac{1-\alpha}{\alpha} w_t$.

But how many units will the monopolist sell or, in other words, what is $X(i)$? As we have indicated above, given the symmetry of the model, where all varieties face the same demand, we can forget about the i label and write $X(i) = \bar{X} = \frac{Z}{M}$. We can thus write the monopolist's profit at any given point in time:

$$\pi_t = \frac{1-\alpha}{\alpha} \frac{L - Z_M}{M_t} w_t. \tag{6.6}$$

So we now see that profits will be a function of Z_M, but we need to find the present discounted value of the flow of profits, and for that we need the interest rate.

That's where we use the NGM, for which the solution is given by (you guessed it) the Euler equation. We can write the interest rate as a function of the growth rate of consumption, namely (with logarithmic utility),

$$r_t = \frac{\dot{c}_t}{c_t} + \rho. \tag{6.7}$$

But consumption must grow at the same rate of output, since all output is consumed in this model. Hence,

$$r_t = \frac{1-\alpha}{\alpha} BZ_M + \rho, \tag{6.8}$$

which is constant. The present value of profits is thus given by[4]

$$\Pi_t = \frac{\frac{1-\alpha}{\alpha}\frac{L-Z_M}{M_t} w_t}{BZ_M + \rho}. \tag{6.9}$$

The free-entry (zero profit) condition requires that the present discounted value of the flow of profits be equal to the cost of creating an additional variety, which (using (6.2)) is given by $\frac{w_t}{BM_t}$. In sum:

$$\frac{\frac{1-\alpha}{\alpha}\frac{L-Z_M}{M_t} w_t}{BZ_M + \rho} = \frac{w_t}{BM_t}. \tag{6.10}$$

Solving this for Z_M allows us to pin down

$$Z_M = (1-\alpha)L - \frac{\alpha\rho}{B}. \tag{6.11}$$

This gives us, at long last, the endogenous growth rate of output:

$$\frac{\dot{Y}_t}{Y_t} = \frac{(1-\alpha)^2}{\alpha} BL - (1-\alpha)\rho, \tag{6.12}$$

again using the fact that the growth rate of Y is $\frac{1-\alpha}{\alpha} BZ_M$.

An increase in the productivity of innovation (B) would lead to a higher growth rate, and, as before, the same would be true for a decrease in the discount rate. So far, so predictable. More importantly, the model shows *scale effects*, as a higher L leads to higher innovation. The intuition is that scale plays a two-fold role, on the supply and on the demand side for ideas. On the one hand, L affects the number of workers in the R&D sector and, as described in (6.2), this increases the production of new varieties. In short, more people means more ideas, which leads to more growth. But L also affects the demand for final output and, hence, for new varieties. This is why profits also depend on the scale of the economy, as can be seen by substituting (6.11) into (6.9). In short, a larger market size allows for bigger profits, and bigger profits make innovation more attractive. This is fundamentally related to the presence of increasing returns. As per the last chapter: developing ideas (new varieties) is a fixed cost in production, and a larger market allows that fixed cost to be further diluted, thereby increasing profits.

The model also brings to the forefront the role of competition, or lack thereof. Innovation is fueled by monopoly profits obtained by the firms that develop new varieties. There is competition in the entry to innovation, of course, which ultimately brings profits to zero once you account for innovation costs. Still, in the absence of monopoly profits in the production of intermediate inputs, there is no incentive to innovate. This immediately highlights the role of policies related to competition and property rights.

We will return to these central insights of endogenous growth models later in the chapter. (Spoiler alert: things are a bit more subtle than in the basic model...)

6.2 | Modelling innovation in quality ladders

We develop a model of innovation through quality ladders, capturing the creative destruction feature of innovation. Besides a similar role as before for scale and monopoly profits as drivers of technological progress and growth, there is now the possibility of excessive growth. Innovation effort may driven by the possibility of replacing existing monopolies and reaping their profits, even when the social payoff of the innovation is small.

The Schumpeterian approach to modelling innovation is associated with Aghion and Howitt (1990) and Grossman and Helpman (1991). We will follow the latter in our discussion.

The model has a continuum of industries $j \in [0, 1]$. Unlike in the previous model, the number of sectors is now fixed, but each of them produces a good with infinite potential varieties. We will think of these varieties as representing different qualities of the product, ordered in a quality ladder. Let's call $q_m(j)$ the quality m of variety j. The (discrete) jumps in quality have size $\lambda > 1$, which we assume exogenous and common to all products so that $q_m(j) = \lambda q_{m-1}(j)$.

The representative consumer has the following expected utility:

$$u_t = \int_0^\infty e^{-\rho t} \left(\int_0^1 \log(\sum_m \ q_m(j)\ x_m(j,t))dj \right) dt,$$

where ρ is the discount factor, and $x_m(j, t)$ is the quantity of variety j (with quality m) consumed in period t. The consumer derives (log) utility from each of the goods and, within each good, preferences are linear. This means that any two varieties are perfect substitutes, which in turn means that the consumer will allocate all their spending on this good to the variety that provides the lowest quality-adjusted cost. As cost of production will be the same in equilibrium, this entails that only the highest-quality variety will be used. Yet, the consumer has the same preferences across varieties, often referred to as Dixit-Stiglitz preferences. They imply that the consumer will allocate their spending equally across varieties, which will come in handy below when solving the model. We call the term $D = \int_0^1 \log \sum_m \ q_m(j)\ x_m(j, t)dj$ the period demand for goods.

All of this can be summarised as follows: If we denote by $E(t)$ the total amount spent in period t, in all goods put together, the solution to the consumer problem implies

$$x_{mt}(j) = \begin{cases} \frac{E(t)}{p_m(j,t)} & \text{if } \frac{q_m(j)}{p_m(j,t)} = \max\{\frac{q_n(j)}{p_n(j,t)}\}\ \forall n\} \\ \\ 0 & \text{if } \frac{q_m(j)}{p_m(j,t)} \neq \max\{\frac{q_n(j)}{p_n(j,t)}\}\ \forall n\}. \end{cases}$$

In words, you spend the same amount on each good, and within each good, only on the highest-quality variety. We can set $E(t)$ equal to one (namely, we choose aggregate consumption to be the numeraire) for simplicity of notation.

The structure of demand provides a fairly straightforward framework for competition. On the one hand, there is monopolistic competition across industries. Within industry, however, competition is fierce because products are perfect substitutes, and firms engage in Bertrand competition with the lowest (quality-adjusted) price taking the whole market. A useful way to think about this is to assume firms have monopoly rights (say, because of patent protection) over their varieties. Thus, only they can produce their variety, but innovators do know the technology so that if they innovate they do so relative to the state-of-the-art producer. This splits the market between state-of-the-art (leading) firms and follower firms, all trying to develop a new highest-quality variety that allows them to dominate the market.

We assume the production function of quality m in industry j to be such that one unit of labour is required to produce one unit of the good. The cost function is trivially

$$c_m(j) = w\, x_m(j),$$

with w being the wage rate. It follows that the minimum price required to produce is w, and, at this price, profits are driven to zero. If followers' price their product at w, the best response for the leading firm is to charge ever-so-slightly below λw, as consumers would still be willing to pay up to that amount given the quality adjustment. For practical purposes, we assume that price to be equal to λw, and this will be common to all industries. Using $E(t) = 1$, profits are trivially given by

$$\pi(t) = x_m(j,t)p_m(j,t) - c_m(j,t) = 1 - \frac{w}{\lambda w} = 1 - \frac{1}{\lambda} = 1 - \delta.$$

where $\delta = 1/\lambda$.

The innovation process is modelled as follows. Firms invest resources, with intensity i for a period dt, to obtain a probability $i\,dt$ of discovering a new quality for the product, and become the state-of-the-art firm. To produce intensity i we assume the firm needs α units of labour with cost $w\alpha$.

Let us think about the incentives to innovate, for the different types of firms. First, note that the state-of-the-art firm has no incentive to innovate. Why so? Intuitively, by investing in R&D the firm has a probability of a quality jump that allows it to set its price at $\lambda^2 w$. This corresponds to an increase in profit of $\frac{\lambda-1}{\lambda^2}$. However, this is smaller than the increase in benefits for followers, for whom profits move from zero to (1-1/λ). In equilibrium, the cost of resources is such that only followers will be able to finance the cost of investment, as they outcompete the state-of-the-art firm for resources, and thus make the cost of capital too high for the latter to turn an expected profit. (Do we really think that leading firms do not invest in innovation? We will return to that later on.)

How about followers? If a follower is successful in developing a better variety, it will obtain a flow of profits in the future. We will denote the present discounted value for the firm as V, which of course will need to consider the fact that the firm will eventually lose its edge because of future innovations. So the firm will invest in R&D if the expected value of innovation is bigger than the cost, that is if $Vidt \geq w\alpha idt$ or $V \geq w\alpha$. In an equilibrium with free entry, we must have $V = w\alpha$.

But what is V, that is, the value of this equity? In equilibrium, it is easy to see that

$$V = \frac{(1-\delta)}{i+\rho}. \tag{6.13}$$

The value of the firm is the discounted value of profits, as usual. But here the discounting has two components: the familiar discount rate, capturing time preferences, and the rate at which innovation may displace this producer.[5]

The final equation that closes the model is the labour market condition. Similar to the model in the previous section, equilibrium in the labour market requires

$$\alpha i + \frac{\delta}{w} = L, \tag{6.14}$$

which means that the labour demand in R&D (αi) plus in production ($\frac{\delta}{w}$) equals the supply of labour.[6]

Equations (6.13) and (6.14), plus the condition that $V = w\alpha$, allow us to solve for the rate of innovation

$$i = (1 - \delta)\frac{L}{\alpha} - \delta\rho. \tag{6.15}$$

Note that innovation is larger the more patient people are, since innovation is something that pays off in the future. It also increases in how efficient the innovation process is – both in terms of the jump it produces and of the cost it takes to obtain the breakthrough. Finally, we once again see scale effects: larger L means larger incentives to innovation. In other words, larger markets foster innovation for the same reasons as in the product-variety model from the previous section . Even though the process of innovation is discrete, by the law of large numbers the process smooths out in the aggregate. This means that the growth rate of consumption is $g = i \ \log\lambda$, which is the growth rate delivered by the model.

What are the implications for welfare? We know from our previous discussion that, in this world with increasing returns and monopolistic behaviour, there can be a wedge between social optimum and market outcomes. But how does that play out here? To answer this question, we can distinguish three effects. First, there is the effect of innovation on consumption, which we can call the *consumer surplus* effect: more innovation produces more quality, which makes consumption cheaper. Second, there is an effect on future innovators, which we can call the *intertemporal spillover* effect: future innovations will occur relative to existing technology, so, by moving the technological frontier, innovation generates additional future benefits. There is, however, a third effect that is negative on current producers that become obsolete, and whose profits evaporate as a result. We call this the *business stealing* effect.

When we put all of this together, a surprising result emerges: the model can deliver a rate of innovation (and growth) that is higher than the social optimum. To see this, imagine that λ is very close to 1, but still larger, such that $\delta = 1 - \nu$ for a small but positive ν. In other words, there is very little social benefit to innovation. However, followers still benefit from innovation because displacing the incumbent (even by a tiny sliver of a margin) gives them monopoly profits. From (6.15) it is clear that, for any given ν, L (and hence profits) can be large enough that we will have innovation although the social value is essentially not there. The divergence between the social and private value, because of the business stealing effect, is what delivers this result.

6.3 | Policy implications

We show how endogenous growth models allow us to think about many policy issues, such as imitation, competition, and market size.

As it turns out, these models of technological change enable us to study a whole host of policy issues as they affect economic growth. Let us consider issues related to distance to the technological frontier, competition policy, and scale effects.

6.3.1 | Distance to the technological frontier and innovation

Put yourself in the role of a policy-maker trying to think about how to foster technological progress and growth. Our models have focused on (cutting-edge) innovation, but there is another way of improving technology: just copy what others have developed elsewhere. What are the implications of that possibility for policy?

To organise our ideas, let us consider a reduced-form, discrete-time setting that captures the flavour of a Schumpeterian model, following Aghion and Howitt (2006). We have an economy with many sectors, indexed by i, each of which has a technology described by

$$Y_{it} = A_{it}^{1-\alpha} K_{it}^{\alpha}, \tag{6.16}$$

where A_{it} is the productivity attained by the most recent technology in industry i at time t, and K_{it} is the amount of capital invested in that sector. If we assume that all sectors are identical *ex ante*, aggregate output (which is the sum of Y_{it}'s) will be given by

$$Y_t = A_t^{1-\alpha} K_t^{\alpha}, \tag{6.17}$$

where A_t is the unweighted sum of A_{it}'s.[7] The Solow model tells us that the long-run growth rate of this economy will be given by the growth rate of A_t. But how is it determined? Following the ideas in the previous section, we assume that, in each sector, only the producer with the most productive technology will be able to stay in business. Now assume that a successful innovator in sector i improves the parameter A_{it}; they will thus be able to displace the previous innovator and become a monopolist in that sector, until another innovator comes along to displace them. This is the creative destruction we have examined.

Now consider a given sector in a given country. A technological improvement in this context can be a new cutting-edge technology that improves on the existing knowledge available in the global economy. Or, more humbly, it can be the adoption of a best practice that is already available somewhere else in the globe. We will distinguish between these two cases by calling them leading-edge and implementation innovation, respectively. As before, leading-edge innovation implies that the innovator obtains a new productivity parameter that is a multiple λ of the previous technology in use in that sector. Implementation, in contrast, implies catching up to a global technology frontier, described by $\bar{A}_t$. We denote μ_n and μ_m the frequency with which leading-edge and implementation innovations take place in that country, as a reduced-form approach of capturing the mechanics from the previous section.

It follows that the change in aggregate productivity will be given by

$$A_{t+1} - A_t = \mu_n \lambda A_t + \mu_m \bar{A}_t + (1 - \mu_n - \mu_m)A_t - A_t = \mu_n(\lambda - 1)A_t + \mu_m(\bar{A}_t - A_t). \tag{6.18}$$

The growth rate will be

$$g = \frac{A_{t+1} - A_t}{A_t} = \mu_n(\lambda - 1) + \mu_m(a_t - 1), \tag{6.19}$$

where $a_t \equiv \frac{\bar{A}_t}{A_t}$ measures the country's average distance to the global technological frontier.

Here's the crucial insight from this simple framework: growth depends on how close the country is to the technological frontier. Given a certain frequency of innovations, being far from the frontier will lead to faster growth since there is room for greater jumps in productivity – the "advantages of backwardness", so to speak, benefit the imitators. The distance to the frontier also affects the mix of innovation that is more growth-enhancing. A country that is far from the frontier will be better off investing more in implementation than another country that is closer to the frontier. This has far-reaching consequences in terms of growth policy. The policies and institutions that foster leading-edge innovation need not be the same as those that foster implementation (see Acemoglu et al. 2006). For instance, think about investment in primary education versus investment in tertiary education, or the role of intellectual property rights protection.[8]

6.3.2 | Competition and innovation

We have seen in the previous sections that the incentive to innovate depends on firms' ability to keep the profits generated by innovation, as captured by the monopoly power innovators acquire. As we pointed out, this formalises an important message regarding the role of monopolies. While monopolies are inefficient in a static context, they are crucial for economic growth.[9] This tradeoff is precisely what lies behind intellectual property rights and the patent system, as had already been noted by Thomas Jefferson in the late 1700s.[10] But is competition always inimical to growth?

The modern Schumpeterian view is more subtle than that. Aghion and coauthors have shown that the relationship between innovation and competition is more complex. The key is that, in addition to this *appropriability* effect, there is also an *escape competition* effect. Increased competition may lead to a greater incentive to innovate as firms will try to move ahead and reap some monopoly profits. In other words, while competition decreases the monopoly rents enjoyed by an innovator, it may decrease the profits of a non-innovator by even more. The overall effect of competition on innovation will critically depend on the nature of where the firms are relative to the frontier. In sectors where competition is neck-and-neck, the escape competition effect is strong. However, if firms are far behind, competition discourages innovation because there is little profit to be made from catching up with the leaders. (Note that this escape competition effect can justify innovation by the leading firms, unlike in the most basic model. In other words, innovation is not simply done by outsiders!)

A similar effect emerges as a result of competition by firms that did not exist previously, namely entrants. We can see that in a simple extension of the reduced-form model above, now focusing on leading-edge innovation. Assume the incumbent monopolist in sector i earns profits equal to

$$\pi_{it} = \gamma A_{it}.$$

In every sector the probability of a potential entrant appearing is p, which is also our measure of entry threat. We focus on technologically advanced entry. Accordingly, each potential entrant arrives with the leading-edge technology parameter $\bar{A}_t$, which grows by the factor λ with certainty each period. If the incumbent is also on the leading edge, with $A_{it} = \bar{A}_t$, then we assume he can use a first-mover advantage to block entry and retain his monopoly. But if he is behind the leading edge, with $A_{it} < \bar{A}_t$, then entry will occur, Bertrand competition will ensue, and the technologically-dominated incumbent will be eliminated and replaced by the entrant.

The effect of entry threat on incumbent innovation will depend on the marginal benefit v_{it}, which the incumbent expects to receive from an innovation. Consider first an incumbent who was on the frontier last period. If they innovate then they will remain on the frontier, and hence will be immune to entry. Their profit will then be $\gamma \bar{A}_t$. If they fail to innovate then with probability p they will be eliminated by entry and earn zero profit, while, with probability $1 - p$, they will survive as the incumbent earning a profit of $\gamma \bar{A}_{t-1}$. The expected marginal benefit of an innovation to this firm is the difference between the profit they will earn with certainty if they innovate and the expected profit they will earn if not:

$$v_{it} = [\lambda - (1-p)]\gamma \bar{A}_{t-1}.$$

Since v_{it} depends positively on the entry threat p, an increase in entry threat will induce this incumbent to spend more on innovating and, hence, to innovate with a larger probability. Intuitively, a firm close to the frontier responds to increased entry threat by innovating more in order to escape the threat.

Next consider an incumbent who was behind the frontier last period, and who will therefore remain behind the frontier even if they manage to innovate, since the frontier will also advance by the factor λ. For this firm, profits will be zero if entry occurs, whether they innovate or not, because they cannot catch up with the frontier. Thus their expected marginal benefit of an innovation will be

$$v_{it} = (1-p)(\lambda - 1)\gamma A_{i,t-1}.$$

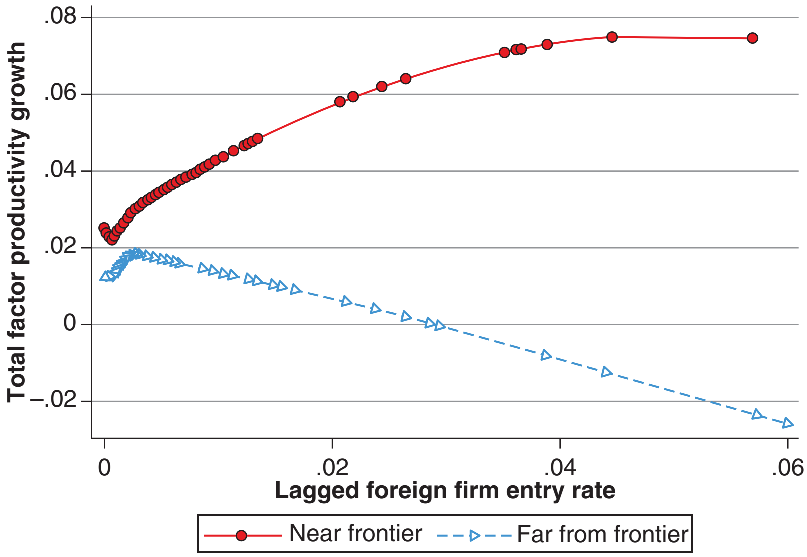
The expected benefit is thus a profit gain that will be realised with probability $(1-p)$, the probability that no potential entrant shows up. Since in this case v_{it} depends negatively on the entry threat p, therefore an increase in entry threat will induce the firm to spend less on innovating. Intuitively, the firm that starts far behind the frontier is discouraged from innovating by an increased entry threat because they are unable to prevent the entrant from destroying the value of their innovation if one happens to show up.

The theory generates the following predictions:

1. Entry and entry threat enhance innovation and productivity growth among incumbents in sectors or countries that are initially close to the technological frontier, as the escape entry effect dominates in that case.
2. Entry and entry threat reduce innovation and productivity growth among incumbents in sectors or countries that are far below the frontier, as the discouragement effect dominates in that case.
3. Entry and entry threat enhance average productivity growth among incumbent firms when the threat has exceeded some threshold, but reduce average productivity growth among incumbents below that threshold. This is because as the probability p measuring the threat approaches unity, then almost all incumbents will be on the frontier, having either innovated or entered last period, and firms near the frontier respond to a further increase in p by innovating more frequently.
4. Entry (and therefore, turnover) is growth-enhancing overall in the short run, because even in those sectors where incumbent innovation is discouraged by the threat of entry, the entrants themselves will raise productivity by implementing a frontier technology.

Figure 6.1, taken from Aghion et al. (2009), provides empirical support for the claim. The graph shows data for UK industries at the four-digit level. Firms are split as those close to the frontier and those away from the frontier (below the sample median for that industry). The level of competition is measured by the rate of foreign firm entry which is measured in the horizontal axis. The vertical axis shows subsequent productivity growth for domestic incumbents. As can be seen, close to the frontier entry accelerates growth. Further away it tends to slow it down.

Figure 6.1 Entry effects, from Aghion et al. (2009)



Interest groups as barriers to innovation

There's another way in which monopolies can affect innovation. Imagine that monopolists use some of their profits to actually block the entry of new firms with better technologies – say, by paying off regulators to bar such entry. In fact, it may be a better deal than trying to come up with innovations! If that's the case, then monopoly profits may actually facilitate the imposition of these barriers, by giving these monopolists more resources to invest in erecting barriers.

Monopolies are particularly dangerous in this regard, because they tend to be better able to act on behalf of their interests. Mancur Olson's *The Logic of Collective Action* Olson (2009) argues that policy is a recurrent conflict between the objectives of concentrated interest groups and those of the general public, for which benefits and costs are typically diffused. According to Olson, the general public has less ability to organise collective action because each actor has less at stake, at least relative to concentrated interest groups, which thus have the upper hand when designing policy. In short, monopolies have an advantage in organising and influencing policy. One implication of this logic is that the seeds of the decline of an economy are contained in its early rise: innovation generates rents that help the development of special-interest lobbies that can then block innovation. This is the argument raised by Mancur Olson, again, in his 1983 book *The Rise and Decline of Nations* Olson (1983).

More recently, the theme of incumbents blocking innovation and development has been taken up by other authors. Parente and Prescott (1999) develop a model capturing this idea, and argue that the effects can be quantitatively large. Restricting the model so that it is consistent with a number of observations between rich and poor countries, they find that eliminating monopoly rights would increase GDP by roughly a factor of 3! Similarly, in their the popular book *Why Nations Fail*, Acemoglu and Robinson (2012) argue, with evidence from a tour de force through history and across every continent, that the typical outcome is that countries fail to develop because incumbents block innovation and disruption.

6.3.3 | Scale effects

The models of the previous section delivered a specific but fundamental result: *scale effects*. In other words, they predict that the growth rates will increase with the size of the population or markets. Intuitively, as we discussed, there are two sides to this coin. On the supply side, if growth depends on ideas, and ideas are produced by people, having more people means having more ideas. On the demand side, ideas are a fixed cost – once you produce a blueprint for a flying car, you can produce an arbitrary amount of flying cars using the same blueprint – and having a larger market enables one to further dilute that fixed cost.

The big question is: do the data support that prediction? Kremer (1993) argues that over the (very) long run of history the predictions of a model with scale effects are verified. He does so by considering what scale effects imply for population growth, which is determined endogenously in his model: population growth is increasing in population. He goes on to test this by checking that, using data from 1,000,000 B.C. to 1990, it does seem to be the case that population growth increases with population size. He also shows that, comparing regions that are isolated from each other (e.g. the continents over pre-modern history), those with greater population displayed faster technological progress.

This suggests that scale effects are present on a global scale, but that remains controversial. For instance, Jones (1995) argues that the data does not support the function in (6.2). For example, the number of scientists involved in R&D grew manifold in the post-World War II period without an increase in the rate of productivity growth. Instead, he argues that the evidence backs a modified version of the innovation production function, in which we would adapt (6.2) to look like this:

$$\frac{\dot{M}}{M} = BZ_M M^{-\beta}, \tag{6.20}$$

with $\beta > 0$. This means that ideas have a diminishing return as you need more people to generate the same rate of innovation. In a world like this, research may deliver a constant rate of innovation (such as the so-called Moore's law on the evolution of the processing capability of computers), but only due to substantially more resources devoted to the activity. This model leads to growth without scale effects, which Jones (1995) refers to as semi-endogenous growth.

It is also worth thinking about what scale effects mean for individual countries. Even if there are scale effects for the global economy, it seems quite obvious that they aren't really there for individual countries: it's not as if Denmark has grown that much slower than the U.S., relative to the enormous difference in size of the two economies. This can be for two reasons. First, countries are not fully isolated from each other, so the benefits of scale leak across borders. Put simply, Danish firms can have access to the U.S. market (and beyond) via trade. This immediately generates a potential connection between trade policy and growth, operating via scale effects. A second reason, on the flip-side, is that countries are not fully integrated domestically, i.e. there are internal barriers to trade that prevent countries from benefiting from their size. A paper by Ramondo et al. (2016) investigates the two possibilities, calibrating a model where countries are divided into regions, and find that the second point is a lot more important in explaining why the Denmarks of the world aren't a lot poorer than the Indias and Chinas and U.S.

6.4 | The future of growth

Are we headed to unprecedented growth? Or to stagnation?

The forces highlighted in these models of innovation, and their policy implications, have huge consequences for what we think will happen in the future when it comes to economic growth. There is a case for optimism, but also for its opposite.

Consider the first. If scale effects are present on a global scale, then as the world gets bigger growth will be faster, not the other way around. To see this, it is worth looking at the Kremer (1993) model in more detail, in a slightly simplified version. Consider the production function in

$$Y = Ap^{\alpha}T^{1-\alpha} = Ap^{\alpha}, \tag{6.21}$$

where p is population and T is land which is available in fixed supply which, for simplicity, we will assume is equal to 1. We can rewrite it as

$$y = Ap^{\alpha-1}. \tag{6.22}$$

The population dynamics have a Malthusian feature in the sense that they revert to a steady state that is sustainable given the technology. In other words, population adjusts to technology so that output per capita remains at subsistence level; as in the Malthusian framework, all productivity gains translate into a larger population, not into higher standards of living. (This is usually thought of as a good description of the pre-industrial era, as we will discuss in detail in Chapter 10.)

$$p = \left(\frac{\bar{y}}{A}\right)^{\frac{1}{\alpha-1}}. \tag{6.23}$$

Critically, the scale effects come into the picture via the assumption that

$$\frac{\dot{A}}{A} = pg. \tag{6.24}$$

i.e. the rate of technological progress is a function of world population, along the lines of the endogenous growth models we have seen. We can now solve for the dynamics of population, using (6.23) and then (6.24):

$$\ln p = \left(\frac{1}{\alpha-1}\right)[\ln \bar{y} - \ln A], \tag{6.25}$$

$$\frac{\dot{p}}{p} = -\left(\frac{1}{\alpha-1}\right)\frac{\dot{A}}{A} = \frac{1}{1-\alpha}\frac{\dot{A}}{A} = \frac{1}{1-\alpha}pg, \tag{6.26}$$

$$\frac{\dot{p}}{p} = \left(\frac{g}{1-\alpha}\right)p. \tag{6.27}$$

In other words, population growth is increasing in population – which means that growth is explosive!

If true, this has enormous consequences for what we would expect growth to be in the future. For instance, if we think that both China and India have recently become much more deeply integrated into the world economy, can you imagine how many ideas these billions of people can come up with?

Can you fathom how much money there is to be made in developing new ideas and selling the resulting output to these billions of people? China and India's integration into the world economy is an added boost to growth prospects on a global scale. In fact, as growth accelerates we may reach a point where machines take over in the accumulation of knowledge, making growth explode without bounds. (You may have heard of this being described as the singularity point.[11])

Yet others have argued that, to the contrary, we are looking at a future of stagnation. Gordon (2018) argues that technological progress has so far relied on three main waves of innovation: the first industrial revolution (steam engine, cotton spinning, and railroads), the second industrial revolution (electricity, internal combustion engine, and running water), and the third industrial revolution (computers and the internet). He argues that the fruit of those waves has been reaped, and mentions a number of factors that may lead to lower future growth:

- The end of the demographic dividend. The process of absorption of women in the labour forces has ended, and the decline in birth rates further pushes down the growth of the labour force.
- Growth in education achievements also has been steadily declining as all the population achieves a minimum standard.
- The scope for growth via imitation falls as previous imitators reach the technological frontier.
- Climate change will require a reduction in future growth.

Some of these factors apply to a greater or lesser degree to other countries. China, for example, also faces a population challenge as its population ages at an unheard-of rate, and education levels have universally improved. At the same time, one might argue that the scope for imitation remains huge. Only assuming that all countries below average world income attain average world income in the next 100 years will deliver an extra 2% growth in world GDP for the next 100 years. Time will tell, we suppose.

6.5 | What have we learned?

In this chapter we presented models of technological innovation, but including technology as a factor of production implies increasing returns to scale, meaning that innovation has to be paid for in some way that cannot be simply via its marginal product. We tackled the issue in three steps. First, we modelled innovation as an increase in the complexity of the production function through a larger number of varieties. Second, as a process of improved quality in varieties which displace previous versions, we developed a framework more akin to Schumpeter's idea of creative destruction. We saw that these two versions both highlight the importance of non-competitive behaviour, with monopoly profits driving the incentive to innovation. They also showcase scale effects: bigger market size implies faster innovation and growth because of the supply and demand for new ideas. In addition, the Schumpeterian version highlighted that there can be too much innovation and growth from a social perspective, as some of the incentive to innovate for private firms is simply to steal monopoly rents from incumbents without a counterpart in social welfare.

We then went over a number of policy issues, through the lens of the models of endogenous growth based on innovation. We saw that distance to the technological frontier can affect the incentives to innovate or imitate. We also saw that the relationship between competition and growth is more subtle than the basic model may indicate. Competition stimulates innovation for firms close to the frontier, but discourages innovation for firms farther away from the frontier. We then went over the debate on the extent to which scale effects matter in practice, presenting a number of arguments on both directions. Finally, we briefly discussed an ongoing debate between those who believe growth will falter and those who think that growth will accelerate.

6.6 | What next?

Acemoglu's (2009) textbook on economic growth provides further details and nuances on the issues discussed here. You can also follow Acemoglu's more recent work on automation. How would the world look if growth accelerates and, for example, robots become ubiquitous? Will this lead to pervasive unemployment? Will this lead to increased income inequality? This has been explored in Acemoglu and Restrepo (2017) and Acemoglu and Restrepo (2020).

In terms of innovation models, an excellent source is the classic Grossman and Helpman (1991) book. In our description of their models we have focused on the steady states, whereas there you will find a full description of the dynamics of the models discussed here. They also go into a lot more detail on the links between trade and economic growth, well beyond our discussion on market size and protection. Similarly, the book by Aghion and Howitt (2008) is a great starting point for further exploration of the Schumpeterian approach that the authors pioneered, and especially on the subtle interplay between competition and innovation. A more recent book by Aghion et al. (2021) also covers and develops these ideas in highly accessible fashion.

If you want more on the debate on the future of growth, the book by Gordon (2017) is a good starting point. That discussion is also the bread-and-butter of futurologists, among which Harari (2018) is a good example. It is interesting to read these books through the lenses of the endogenous growth models we have seen here.

Notes

1 You may also notice the new exponent $\frac{1}{\alpha}$, which will afford notational simplicity as we pursue the algebra.

2 This happens to be a property of the equilibrium of this model, and not an assumption, but we will simply impose it here for simplicity.

3 To see this, note that differentiating the term $\left[\int_0^M X(i)^\alpha \, di\right]^{\frac{1}{\alpha}}$ with respect to $X(i)$ yields $X(i)^{\alpha-1}$ $\left[\int_0^M X(i)^\alpha \, di\right]^{\frac{1-\alpha}{\alpha}}$, and $\left[\int_0^M X(i)^\alpha \, di\right]^{\frac{1-\alpha}{\alpha}} = 1$ because of our normalisation to unit output.

4 Why is the denominator $BZ_M + \rho$ the appropriate discount rate by which to divide π_t? If π_t were constant, obtaining the present value of profits would require simply dividing it by the (constant) interest rate. But π_t is not constant: it grows at the rate at which $\frac{w_t}{M_t}$ grows. Since wages must in equilibrium grow at the rate of output, it follows that $\frac{w_t}{M_t}$ grows at the growth rate of output minus the growth rate of M_t: $\frac{1-2\alpha}{\alpha} BZ_M$. Subtracting this from the interest rate gives us the appropriate discount rate: $BZ_M + \rho$.

5 We can use the consumer discount rate ρ because we assume firms are held in a diversified portfolio and there is no idiosyncratic risk.

6 Labour demand in production follows from this: each sector uses one unit of labor per unit of the good being produced. With total expenditure normalized to one, it follows that they sell $x = \frac{1}{p} = \frac{1}{lambdaw} = \frac{\delta}{w}$ units each, which integrated between 0 and 1, for all sectors, yields the result.

[7] Not quite the sum, but close enough. For those of you who are more mathematically inclined: define $x_{it} \equiv \log X_{it}$ for any X, then from (6.16) you can write $y_{it} = (1-\alpha)a_{it} + \alpha k_{it}$. Now assume the many sectors in the economy fall in the interval $[0,1]$, and integrate y_{it} over that interval: $\int_0^1 y_{it} di = (1-\alpha)\int_0^1 a_{it} + \alpha \int_0^1 k_{it} \Rightarrow y_t = (1-\alpha)a_t + \alpha k_t$. Define $Y_t \equiv \exp(y_t)$, and (6.17) follows. In sum, (6.17) essentially defines $X_t = \exp\left((\int_0^1 \log X_{it} di)\right)$ for any variable X. These are all monotonic transformations, so we are fine.

[8] Williams (2013) shows an interesting piece of evidence: gene sequences subject to IP protection by private firm Celera witnessed less subsequent scientific research and product development, relative to those sequenced by the public Human Genome Project.

[9] This is an insight that Schumpeter himself had pioneered in his book *Capitalism, Socialism and Democracy*, back in 1942. See Schumpeter (1942).

[10] See https://www.monticello.org/site/research-and-collections/patents.

[11] As in $\frac{1}{x}$ when it approaches 0.

References

Acemoglu, D. (2009). *Introduction to modern economic growth*. Princeton University Press.

Acemoglu, D., Aghion, P., & Zilibotti, F. (2006). Distance to frontier, selection, and economic growth. *Journal of the European Economic Association*, *4*(1), 37–74.

Acemoglu, D. & Restrepo, P. (2017). Secular stagnation? The effect of aging on economic growth in the age of automation. *American Economic Review*, *107*(5), 174–79.

Acemoglu, D. & Restrepo, P. (2020). Robots and jobs: Evidence from U.S. labor markets. *Journal of Political Economy*, *128*(6), 2188–2244.

Acemoglu, D. & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Currency.

Aghion, P., Blundell, R., Griffith, R., Howitt, P., & Prantl, S. (2009). The effects of entry on incumbent innovation and productivity. *The Review of Economics and Statistics*, *91*(1), 20–32.

Aghion, P. & Howitt, P. (1990). *A model of growth through creative destruction*. National Bureau of Economic Research.

Aghion, P. & Howitt, P. (2006). Appropriate growth policy: A unifying framework. *Journal of the European Economic Association*, *4*(2-3), 269–314.

Aghion, P. & Howitt, P. W. (2008). *The economics of growth*. MIT Press.

Gordon, R. J. (2017). *The rise and fall of American growth: The U.S. standard of living since the Civil War*. Princeton University Press.

Gordon, R. J. (2018). *Why has economic growth slowed when innovation appears to be accelerating?* National Bureau of Economic Research.

Grossman, G. M. & Helpman, E. (1991). *Innovation and growth in the global economy*. MIT Press.

Harari, Y. N. (2018). *21 lessons for the 21st century*. Random House.

Jones, C. I. (1995). R & d-based models of economic growth. *Journal of Political Economy*, *103*(4), 759–784.

Kremer, M. (1993). Population growth and technological change: One million B.C. to 1990. *The Quarterly Journal of Economics*, *108*(3), 681–716.

Olson, M. (1983). *The rise and decline of nations: Economic growth, stagflation, and social rigidities*. Yale University Press.

Olson, M. (2009). *The logic of collective action: Public goods and the theory of groups, second printing with a new preface and appendix*. Harvard University Press.

Parente, S. L. & Prescott, E. C. (1999). Monopoly rights: A barrier to riches. *American Economic Review*, *89*(5), 1216–1233.

Ramondo, N., Rodríguez-Clare, A., & Saborío-Rodríguez, M. (2016). Trade, domestic frictions, and scale effects. *American Economic Review*, *106*(10), 3159–84.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, *98*(5, Part 2), S71–S102.

Schumpeter, J. A. (1942). Capitalism. *Socialism and democracy*, *3*, 167.

Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy*, *121*(1), 1–27.

CHAPTER 7

Proximate and fundamental causes of growth

Now let's talk a little bit about what the data say regarding economic growth. There is a very long line of research trying to empirically assess the determinants of growth – an area that is still very vibrant. In order to organise what this literature has to say, it is useful to start by distinguishing between what Acemoglu (2009) calls *proximate* and *fundamental* causes of economic growth. If we think of any generic production function $Y = F(\mathbf{X}, A)$, where $\mathbf{X}$ is a vector of inputs (capital, labour, human capital) and A captures productivity, we can attribute any increase in output to an increase in $\mathbf{X}$ or A. In that sense, the accumulation of physical capital, human capital, or technological progress generates growth, but we still want to learn why different societies choose different accumulation paths. We can thus think of these as *proximate* causes, but we want to be able to say something about the *fundamental* causes that determine those choices. Our survey of the empirical literature will address what economists have been able to say about each of those sets of causes.

7.1 | The proximate causes of economic growth

There are three basic empirical tools to assess the importance of proximate causes of growth (factor accumulation, productivity): growth accounting, regression-based approaches, and calibration. We briefly go over the advantages and pitfalls, and the message they deliver. Factor accumulation has significant explanatory power, but in the end productivity matters a lot.

The natural starting point for this investigation is our workhorse, the Neoclassical Growth Model (NGM). The basic question, to which we have already alluded, is: how well does the NGM do in explaining differences in income levels and in growth rates?[1]

Several methods have been devised and used to assess this question, and they can be broadly grouped into three classes: *growth accounting, growth regressions, and calibration*. Let us address each of these.

How to cite this book chapter:
Campante, F., Sturzenegger, F. and Velasco, A. 2021. *Advanced Macroeconomics: An Easy Guide.* Ch. 7. 'Proximate and fundamental causes of growth', pp. 87–112. London: LSE Press.
DOI: https://doi.org/10.31389/lsepress.ame.g

7.1.1 | Growth accounting

This is another founding contribution of Robert Solow to the study of economic growth. Right after publishing his "Contribution to the Theory of Economic Growth" in 1956, he published another article in 1957 (Solow 1957) noting that an aggregate production function such as

$$Y(t) = A(t) F\left(K_t, L_t\right), \tag{7.1}$$

when combined with competitive factor markets, immediately yields a framework that lets us account for the (proximate) sources of economic growth. Take the derivative of the log of the production function with respect to time,

$$\frac{\dot{Y}}{Y} = \frac{\dot{A}}{A} + \frac{AF_K}{Y}\dot{K} + \frac{AF_L}{Y}\dot{L} \Rightarrow$$

$$\frac{\dot{Y}}{Y} = \frac{\dot{A}}{A} + \frac{AF_K K}{Y}\frac{\dot{K}}{K} + \frac{AF_L L}{Y}\frac{\dot{L}}{L} \Rightarrow$$

$$g_Y = g_A + \alpha_K g_K + \alpha_L g_L, \tag{7.2}$$

where g_X is the growth rate of variable X, and $\alpha_X \equiv \frac{AF_X X}{Y}$ is the elasticity of output with respect to factor X. This is an identity, but adding the assumption of competitive factor markets (i.e. factors are paid their marginal productivity) means that α_X is also the share of output that factor X obtains as payment for its services. Equation (7.2) then enables us to estimate the contributions of factor accumulation and technological progress (often referred to as total factor productivity (TFP)) to economic growth.

This is how it works in practice: from national accounts and other data sources, one can estimate the values of g_Y, g_K, g_L, α_K, and α_L; from (7.2) one can then back out the estimate for g_A.[2] (For this reason, g_A is widely referred to as the Solow residual.) Solow actually computed this for the U.S. economy, and reached the conclusion that the bulk of economic growth, about 2/3, could be attributed to the residual. Technological progress, and not factor accumulation, seems to be the key to economic growth.

Now, here is where a caveat is needed: g_A is calculated as a residual, not directly from measures of technological progress. It is the measure of our ignorance![3] More precisely, any underestimate of the increase in K or L (say, because it is hard to adjust for the increased quality of labour input), will result in an overestimate of g_A. As a result, a lot of effort has been devoted to better measure the contribution of the different factors of production.

In any event, this approach has been used over and over again. A particularly famous example was Alwyn Young's research in the early 1990s (1995), where he tried to understand the sources of the fantastic growth performance of the East Asian "tigers", Hong Kong, Singapore, South Korea, and Taiwan.[4] Most observers thought that this meant that they must have achieved amazing rates of technological progress, but Young showed that their pace of factor accumulation had been astonishing. Rising rates of labour force participation (increasing L), skyrocketing rises in investment rates (from 10% of GDP in 1960 to 47% of GDP in 1984, in Singapore, for instance!) (increasing K), and increasing educational achievement (increasing H). Once all of this is accounted for, their Solow residuals were not particularly outliers compared to the rest of the world. (This was particularly the case for Singapore, and not so much for Hong Kong.) Why is this important? Well, we know from the NGM that factor accumulation cannot sustain growth in the long run! This seemed to predict that the tigers' performance would soon hit the snag of decreasing returns. Paul Krugman started to become famous beyond the circles of economics by explicitly predicting as much in a famous article in 1994 (Krugman 1994), which was interpreted by many as having predicted the 1997 East Asian crisis.

Of course, the tigers resumed growing fast soon after that crisis – have they since then picked up with productivity growth?

7.1.2 | Using calibration to explain income differences

We have seen in Chapter 2 that a major issue in growth empirics is to assess the relative importance of factor accumulation and productivity in explaining differences in growth rates and income levels. A different empirical approach to this question is *calibration*, in which differences in productivity are calculated using imputed parameter values that come from microeconomic evidence. As it is closely related to the methodology of growth accounting, we discuss it here. (We will see later, when discussing business cycle fluctuations, that calibration is one of the main tools of macroeconomics, when it comes to evaluating models empirically.)

One of the main contributions in this line of work is a paper by Hall and Jones (1999). In their approach, they consider a Cobb-Douglas production function for country i,

$$Y_i = K_i^{\alpha} \left(A_i H_i\right)^{1-\alpha}, \tag{7.3}$$

where K_i is the stock of physical capital, H_i is the amount of human capital-augmented labour and A_i is a labour-augmenting measure of productivity. If we know α, K_i and H_i, and given that we can observe Y, we can back out productivity A_i:

$$A_i = \frac{Y_i^{\frac{1}{1-\alpha}}}{K_i^{\frac{\alpha}{1-\alpha}} H_i}. \tag{7.4}$$

But how are we to know those?

For human capital-augmented labour, we start by assuming that labour L_i is homogeneous within a country, and each unit of it has been trained with E_i years of schooling. Human capital-augmented labour is given by

$$H_i = e^{\phi(E_i)} L_i. \tag{7.5}$$

The function $\phi(E)$ reflects the efficiency of a unit of labour with E years of schooling relative to one with no schooling $(\phi(0) = 0)$. $\phi'(E)$ is the return to schooling estimated in a Mincerian wage regression (i.e. a regression of log wages on schooling and demographic controls, at the individual level). As such, we can run a Mincerian regression to obtain H_i. (Hall and Jones do so assuming that different types of schooling affect productivity differently.)

How about physical capital? We can compute it from data on past investment, using what is called the perpetual inventory method. If we have a depreciation rate δ, it follows that

$$K_{i,t} = (1-\delta)K_{i,t-1} + I_{i,t-1}. \tag{7.6}$$

It also follows that

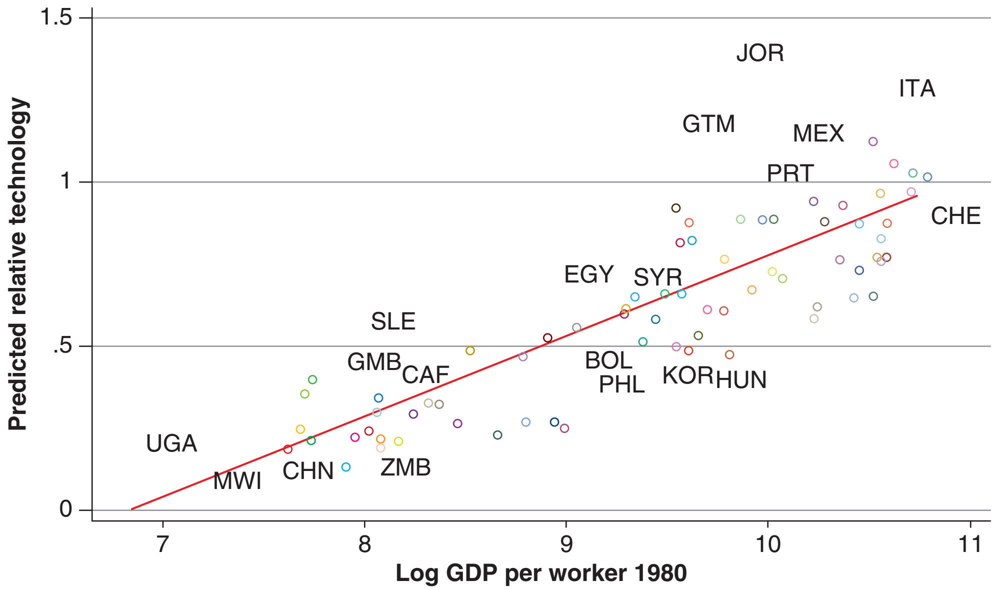
$$K_{i,t} = (1-\delta)^t K_{i,0} + \sum_{s=0}^{t} I_{i,s}(1-\delta)^{t-s-1}. \tag{7.7}$$

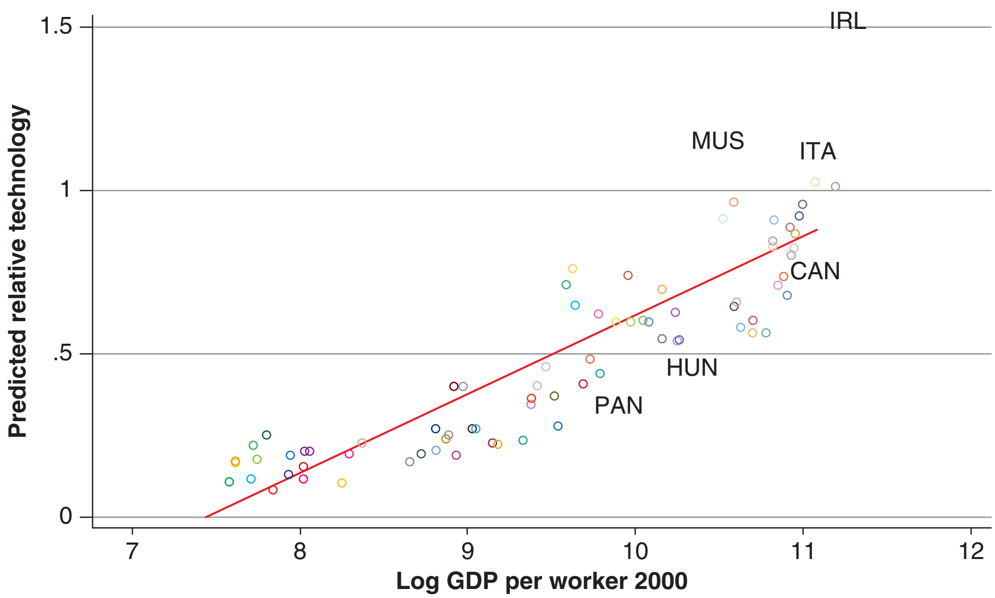
If we have a complete series of investment, we can calculate this for any point in time. (We assume $\delta = 0.06$ for all countries). Since we don't, we assume that, before the start of our data series, investment had been growing at the same rate that we observe in the sample. By doing that, we can compute the $K_{i,0}$ and obtain our value for the capital stock.

How about α? Well, we go for our usual assumption of $\alpha = 1/3$, which is thought of as a reasonable value given the share of capital returns in output as measured by national accounts. This is subject to the caveats we have already discussed, but it is a good starting point.

Since we are interested in cross-country comparisons, we benchmark the data with comparisons to the U.S. series. This comparison can be seen in Figure 7.1, from Acemoglu (2009).

Figure 7.1 Productivity differences, from Acemoglu (2012)





If all countries had the same productivity, and all differences in income were due to differences in factor accumulation, we would see all countries bunched around a value of 1 in the y-axis. This is clearly not the case! Note also that the pattern seems to become stronger over time: we were farther from that benchmark in 2000 than in 1980.

To summarise the message quantitatively, we can do the following exercise. Output per worker in the five countries with the highest levels of output per worker was 31.7 times higher than output per worker in the five lowest countries. Relatively little of this difference was due to physical and human capital:

- Capital intensity per worker contributed a factor of 1.8
- Human capital per worker contributed a factor of 2.2
- Productivity contributed a factor of 8.3!

Hall and Jones associate this big impact of productivity to the role of social capital: the ability of societies to organise their economic activity with more or less costs. For example, a society where theft is prevalent will imply the need to spend resources to protect property; a society full of red tape would require lots of energy in counteracting it, and so on. In short, productivity seems a much bigger concept than just technological efficiency.

However, just as in the regression approaches, calibration also relies on important assumptions. Now, functional forms make a huge difference, both in the production function and in the human capital equation. If we lift the Cobb-Douglas production function or change the technological assumptions in the production of human capital (e.g. assuming externalities), things can change a lot.

7.1.3 | Growth regressions

Another approach to the empirics of economic growth is that of growth regressions – namely, estimating regressions with growth rates as dependent variables. The original contribution was an extremely influential paper by Robert Barro (1991), that established a canonical specification. Generally speaking, the equation to be estimated looks like this:

$$g_{i,t} = \mathbf{X}'_{i,t}\beta + \alpha \log(y_{i,t-1}) + \epsilon_{i,t}, \tag{7.8}$$

where $g_{i,t}$ is the growth rate of country i from period $t-1$ to period t, $\mathbf{X}'_{i,t}$ is a vector of variables that one thinks can affect a country's growth rate, both in steady state (i.e. productivity) and along the transition path, β is a vector of coefficients, $y_{i,t-1}$ is country i's output in the previous period $t-1$, α is a coefficient capturing convergence, and $\epsilon_{i,t}$ is a random term that captures all other factors omitted from the specification.

Following this seminal contribution, innumerable papers were written over the subsequent few years, with a wide range of results. In some one variable was significant; in others, it was not. Eventually, the results were challenged on the basis of their robustness. Levine and Renelt (1991), for example, published a paper in which they argued *no* results were robust. The counterattack was done by a former student and colleague of Barro, Sala-i-Martin (1997), that applied a similar robustness check to all variables used by any author in growth regressions, in his amusingly titled paper, "I Just Ran Two Million Regressions". He concluded that, out of the 59 variables that had shown up as significant somewhere in his survey of the literature, some 22 seem to be robust according to his more lax, or less extreme, criteria (compared to Levine and Renelt's). These include region and religion dummies, political variables (e.g. rule of law), market distortions (e.g. black market premium), investment, and openness.

Leaving aside the issues of robustness, the approach, at least in its basic form, faces other severe challenges, which are of two types, roughly speaking.

1. **Causality** (aka Identification; aka Endogeneity): The variables in $\mathbf{X}_{i,t}$ are typically endogenous, i.e. jointly determined with the growth rate. As you have seen in your courses on econometrics, this introduces bias in our estimates of β, which in turn makes it unlikely that we identify the causal effect of any of those variables (at the end of this chapter, when discussing institutions, we will discuss the solution of this problem suggested by Acemoglu et al. (2001), one of the most creative and influential proposed solutions to this endogeneity problem).
2. **Interpretation**: The economic interpretation of the results might be quite difficult. Suppose that openness truly affects economic growth, but it does so (at least partly) by increasing investment rates; if our specification includes both variables in $\mathbf{X}_{i,t}$, the coefficient on openness will not capture the full extent of its true effect.

Both of these are particularly problematic if we want to investigate the relationship between policies and growth, a point that is illustrated by Dani Rodrik's (2012) critique. Rodrik's point is that if policies are endogenous (and who could argue they are not?) we definitely have a problem. The intuition is as follows. Imagine you want to test whether public banks are associated with higher or lower growth. If you run a regression of growth on, say, share of the financial sector that is run by public banks, you may find a negative coefficient. But is that because public banks are bad for growth? Or is it because politicians resort to public banks when the economy faces a lot of constraints (and thus its growth is bound to be relatively low)?

To see the issue more clearly, consider a setup, from a modified AK model, in which

$$g = (1 - \theta) A - \rho, \tag{7.9}$$

where θ is a distortion. Now consider a policy intervention s, which reduces the distortion, but that has a cost of its own. Then,

$$g(s, \theta, \phi) = (1 - \theta (1 - s)) A - \phi\alpha (s) - \rho. \tag{7.10}$$

The optimal intervention delivers growth as defined by the implicit equation

$$g_s (s^{**}, \theta, \phi) = 0. \tag{7.11}$$

In addition, there is a diversion function of the policy maker $\pi (s)$, with $\pi' (s) > 0$, $\pi'' (s) < 0$, and $\pi' \left(s^P\right) = 0$ with $s^P > s^{**}$. This means that the politicians will use the intervention more than is actually desirable from a social perspective. The politician will want to maximise growth and their own benefit, placing a weight λ on growth. This means solving

$$\max_s u (s, \theta, \phi) = \lambda g (s, \theta, \phi) + \pi (s), \tag{7.12}$$

which from simple optimisation yields the FOC

$$\lambda g_s (s^*, \theta, \phi) + \pi' (s^*) = 0. \tag{7.13}$$

Because we have assumed that $\pi' (s) > 0$, it follows from (7.13) that $g_s (s^*, \theta, \phi) < 0$, and this implies that a reduction in s will increase growth. Does this imply that we should reduce s? Marginally, yes, but not to zero, which is the conclusion that people typically read from growth regressions.

Now, what if we were to run a regression to study the links between policy s and growth? We need to take into account the changes in s that happen when the parameters vary across countries. Consider the effect of changes in the level of distortions θ. Recall that, from (7.10):

$$g_s(s,\theta,\phi) = \theta A - \phi\alpha'(s). \tag{7.14}$$

Replacing in (7.13) and totally differentiating yields

$$d\theta\lambda A + \left[-\lambda\phi\alpha''(s) + \pi''(s^*)\right] ds^* = 0 \tag{7.15}$$

$$\frac{ds^*}{d\theta} = \frac{\overset{(+)}{\lambda A}}{\underbrace{\lambda\phi\alpha''(s) - \pi''(s^*)}_{(+)}} > 0. \tag{7.16}$$

This implies that in an economy with greater inefficiencies we will see a higher level of the policy intervention, as long as politicians care about growth. But growth will suffer with the greater inefficiencies: differentiating (7.10) with respect to θ we have

$$\frac{dg}{d\theta} = -A(1-s^*) + g_s(s^*,\theta,\phi)\frac{ds^*}{d\theta} < 0 \Rightarrow \frac{\frac{dg}{d\theta}}{\frac{ds^*}{d\theta}} < 0. \tag{7.17}$$

The fact that this coefficient is negative means nothing, at least from a policy perspective (remember that it is optimal to increase the policy intervention if the distortion increases).

Because of challenges like these, people later moved to analyse panel growth regressions, which rearrange (7.8) as

$$g_{i,t} = \mathbf{X}'_{i,t}\beta + \alpha\log(y_{i,t-1}) + \delta_i + \mu_t + \epsilon_{i,t}, \tag{7.18}$$

where δ_i and μ_t are country and time fixed effects, respectively. By including country fixed effects, this removes fixed country characteristics that might affect both growth and other independent variables of interest, and thus identifies the effects of such variables out of within-country variation. However, in so doing they might be getting rid of most of the interesting variation, which is across countries, while also increasing the potential bias due to measurement error. Finally, these regressions do not account for time-varying country-specific factors. In sum, they are no panacea.

Convergence

Another vast empirical debate that has taken place within the framework of growth regressions is one to which we have already alluded when discussing the Solow model: convergence. We have talked, very briefly, about what the evidence looks like; let us now get into more detail.

Absolute convergence

As you recall, this is the idea that poorer countries grow faster than richer countries, unconditionally. Convergence is a stark prediction of the NGM, and economists started off by taking this (on second thought, naive) version of it to the data. Baumol (1986) took Maddison's core sample of 16 rich countries over the long run and found

$$\underset{}{\text{growth}} = 5.251 - \underset{(0.075)}{0.749} \text{ initial income}$$

with $R^2 = 0.87$. He thus concluded that there was strong convergence!

However, De Long (1988) suggested a reason why this result was spurious: only successful countries took the effort to construct long historical data series. So the result may be a simple fluke of sample selection bias (another problem is measurement error in initial income that also biases the results in favour of the convergence hypothesis). In fact, broadening the sample of countries beyond Madison's sixteen leads us immediately to reject the hypothesis of convergence. By the way, there has been extensive work on convergence, within countries and there is fairly consistent evidence of absolute convergence for different regions of a country.[5]

Conditional convergence

The literature then moved to discuss the possibility of conditional convergence. This means including in a regression a term for the initial level of GDP, and checking whether the coefficient is negative when controlling for the other factors that determine the steady state of each economy. In other words, we want to look at the coefficient α in (7.8), which we obtain by including the control variables in $\mathbf{X}$. By including those factors in the regression, we partial out the steady state from initial income and measure deviations from this steady state. This, of course, is the convergence that is actually predicted by the NGM.

Barro (1991) and Barro and Sala-i-Martin (1992) found evidence of a negative α coefficient, and we can say that in general the evidence is favourable to conditional convergence. Nevertheless, the same issues that apply to growth regressions in general will be present here as well.

7.1.4 | Explaining cross-country income differences, again

Another regression-based approach to investigate how the NGM fares in explaining the data was pioneered by Mankiw et al. (1992) (MWR hence). Their starting point is playfully announced in the very first sentence: "This paper takes Robert Solow seriously" (p. 407).[6] This means that they focus simply on the factor accumulation determinants that are directly identified by the Solow model as the key proximate factors to explain cross-country income differences, leaving aside the productivity differences. They claim that the NGM (augmented with human capital) does a good job of explaining the existing cross-country differences.

Basic Solow model

There are two inputs, capital and labour, which are paid their marginal products. A Cobb-Douglas production function is assumed

$$Y_t = K_t^{\alpha} \left(A_t L_t\right)^{1-\alpha} \qquad 0 < \alpha < 1. \tag{7.19}$$

L and A are assumed to grow exogenously at rates n and g:

$$\frac{\dot{L}}{L} = n \tag{7.20}$$

$$\frac{\dot{A}}{A} = g. \tag{7.21}$$

The number of effective units of labour $A(t)L(t)$ grows at rate $n + g$.

As usual, we define k as the stock of capital per effective unit of labour $k = \frac{K}{AL}$ and $y = \frac{Y}{AL}$ as the level of output per effective unit of labour.

The model assumes that a constant fraction of output s is invested. The evolution of k is

$$\dot{k}_t = sy_t - \left(n + g + \delta\right) k_t \tag{7.22}$$

or

$$\dot{k}_t = sk_t^{\alpha} - \left(n + g + \delta\right) k_t, \tag{7.23}$$

where δ is the rate of depreciation. The steady state value k^* is

$$k^* = \left[\frac{s}{\left(n + g + \delta\right)}\right]^{\frac{1}{1-\alpha}}. \tag{7.24}$$

Output per capita is

$$\left(\frac{Y_t}{L_t}\right) = K_t^{\alpha} A_t^{1-\alpha} L_t^{-\alpha} = k_t^{\alpha} A_t. \tag{7.25}$$

Substituting (7.24) into (7.25)

$$\left(\frac{Y_t}{L_t}\right) = \left[\frac{s}{\left(n + g + \delta\right)}\right]^{\frac{\alpha}{1-\alpha}} A_t \tag{7.26}$$

and taking logs

$$\log\left(\frac{Y_t}{L_t}\right) = \frac{\alpha}{1-\alpha}\log(s) - \frac{\alpha}{1-\alpha}\log\left(n + g + \delta\right) + \log A(0) + gt. \tag{7.27}$$

MRW assume that g (representing advancement of knowledge) and δ do not vary across countries, but A reflects not only technology but also resource endowments. It thus differs across countries as in

$$\log A(0) = a + \epsilon, \tag{7.28}$$

where a is a constant and ϵ is a country-specific shock. So we have

$$\log\left(\frac{Y}{L}\right) = a + \frac{\alpha}{1-\alpha}\log(s) - \frac{\alpha}{1-\alpha}\log\left(n + g + \delta\right) + \epsilon \tag{7.29}$$

We assume s and n are not correlated with ϵ. (What do you think of this assumption?) Since it is usually assumed that the capital share is $\alpha \cong \frac{1}{3}$, the model predicts an elasticity of income per capita with respect to the saving rate $\frac{\alpha}{1-\alpha} \cong \frac{1}{2}$ and an elasticity with respect to $n+g+\delta$ of approximately -0.5.

Table 7.1 Estimates of the basic Solow model

	Update Log GDP per Capita		
	MRW 1985	Acemoglu 2000	Update 2017
$\log(s_k)$	1.42*** (.14)	1.22*** (.13)	.96* (.48)
$\log(n+g+\delta)$	−1.97*** (.56)	−1.59*** (.36)	−1.48*** (.21)
Implied α	.59	.55	.49
Adjusted R^2	.59	.49	.49

Note: *p<0.1; **p<0.05; ***p<0.01

What do the data say?

With data from the real national accounts constructed by Summers and Heston (1988) for the period 1960-1985, they run (7.29), using ordinary least squares (OLS) for all countries for which data are available minus countries where oil production is the dominant industry.

We reproduce their results in Table 7.1, to which we add an update by Acemoglu (2009), and one of our own more than 30 years after the original contribution. In all three cases, aspects of the results support the Solow model:

1. Signs of the coefficients on saving and population growth are OK.
2. Equality of the coefficients for $\log(s)$ and $-\log(n+g+\delta)$ is not rejected.
3. A high percentage of the variance is explained (see R^2 in the table).

But the estimate for α contradicts the prediction that $\alpha = 1/3$. While the implicit value of α seems to be falling, in each update it is still around or above .5. Some would have said it is OK (remember our discussion in Chapter 2), but for MRW it was not.

Introducing human capital

MRW go on to consider the implications of considering the role of human capital. Let us now recall the augmented Solow model that we saw in Chapter 5. The production function is now

$$Y_t = K_t^{\alpha} H_t^{\beta} \left(A_t L_t\right)^{1-\alpha-\beta}, \tag{7.30}$$

where H is the stock of human capital. If s_k is the fraction of income invested in physical capital and s_h the fraction invested in human capital, the evolution of k and h are determined by

$$\dot{k}_t = s_k y_t - \left(n + g + \delta\right) k_t \tag{7.31}$$

$$\dot{h}_t = s_h y_t - \left(n + g + \delta\right) h_t, \tag{7.32}$$

where k, h and y are quantities per effective unit of labour.

It is assumed that $\alpha + \beta < 1$, so that there are decreasing returns to all capital and we have a steady state for the model. The steady-state level for k and h are

$$k^* = \left[\frac{s_k^{1-\beta} s_h^{\beta}}{(n+g+\delta)}\right]^{\frac{1}{1-\alpha-\beta}} \tag{7.33}$$

$$h^* = \left[\frac{s_k^{\alpha} s_h^{1-\alpha}}{(n+g+\delta)}\right]^{\frac{1}{1-\alpha-\beta}}. \tag{7.34}$$

Substituting (7.33) and (7.34) into the production function and taking logs, income per capita is

$$\begin{aligned} \log\left(\frac{Y_t}{L_t}\right) = \frac{\alpha}{1-\alpha-\beta}\log\left(s_k\right) + \frac{\beta}{1-\alpha-\beta}\log\left(s_h\right) \\ - \frac{\alpha+\beta}{1-\alpha-\beta}\log\left(n+g+\delta\right) + \log A(0) + gt. \end{aligned} \tag{7.35}$$

To implement the model, investment in human capital is restricted to education. They construct a SCHOOL variable that measures the percentage of the working age population that is in secondary school, and use it as a proxy for human capital accumulation s_h.

The results are shown in Table 7.2. It turns out that now 78% of the variation is explained, and the numbers seem to match: $\hat{\alpha} \cong 0.3$, $\hat{\beta} \cong 0.3$. (For the updated data we have a slightly lower R^2 and a higher $\hat{\beta}$ indicating an increasing role of human capital, in line with what we found in the previous section.)

Table 7.2 Estimates of the augmented Solow model

	Update		
	Log GDP per Capita		
	MRW 1985	Acemoglu 2000	Update 2017
$\log(s_k)$	$.69^{***}$ (.13)	$.96^{***}$ (.13)	.71 (.44)
$\log(n+g+\delta)$	-1.73^{***} (.41)	-1.06^{***} (.33)	-1.43^{***} (.19)
$\log(s_h)$	$.66^{***}$ (.07)	$.70^{***}$ (.13)	1.69^{***} (.43)
Implied α	.30	.36	.28
Implied β	.28	.26	.33
Adjusted R^2	.78	.60	.59

Note: *p<0.1; **p<0.05; ***p<0.01

Challenges

The first difficulty with this approach is: is it really OK to use OLS? Consistency of OLS estimates requires that the error term be orthogonal to the other variables. But that error term includes technology differences, and are these really uncorrelated with the accumulation of physical and human capital? If not, the omitted variable bias (and reverse causality) would mean that the estimates for the effects of physical and human capital accumulation (and for the R^2) are biased upwards, and the NGM doesn't do as good a job as MRW think, when it comes to explaining cross-country income differences. This is the very same difficulty that arises from the growth regressions approach – not surprising, since the econometric underpinnings are very much similar.

A second difficulty has to do with the measure of human capital: is it really a good one? The microeconometric evidence suggests that the variation in average years of schooling across countries that we see in the data is not compatible with the estimate $\hat{\beta}$ obtained by MRW.

7.1.5 | Summing up

We have seen many different empirical approaches, and their limitations. Both in terms of explaining differences in growth and in income levels at the cross-country level, there is a lot of debate on the extent to which the NGM can do the job.

It does seem that the consensus in the literature today is that productivity differences are crucial for understanding cross-country differences in economic performance. (A paper by Acemoglu and Dell (2010) makes the point that productivity differences are crucial for *within*-country differences as well.) This means that the endogenous growth models that try to understand technological progress are a central part of understanding those differences.

In the previous chapter we talked about some of the questions surrounding those models, such as the effects of competition and scale, but these models focused on productive technology, that is, how to build a new blueprint or a better variety for a good. The empirical research, as we mentioned above, suggests that productivity differences don't necessarily mean technology in a narrow sense. A country can be less productive because of market or organisational failures, even for a given technology. The reasons for this lower productivity may be manifold, but they lead us into the next set of questions: what explains them? What explains differences in factor accumulation? In other words, what are the *fundamental* causes of economic performance? We turn to this question now.

7.2 | The fundamental causes of economic growth

We go over four types of fundamental explanations for differences in economic performance: luck (multiple equilibria), geography, culture, and institutions.

As North (1990) point out, things like technological progress and factor accumulation "are not causes of growth; *they are growth*" (p.2). The big question is, what in turn causes them? Following Acemoglu (2009), we can classify the main hypotheses into four major groups:

1. *Luck*: Countries that are identical in principle may diverge because small factors lead them to select different equilibria, assuming that multiple equilibria exist.

2. *Geography*: Productivity can be affected by things that are determined by the physical, geographical, and ecological environment: soil quality, presence of natural resources, disease environment, inhospitable climate, etc.
3. *Culture*: Beliefs, values, and preferences affect economic behaviour, and may lead to different patterns of factor accumulation and productivity: work ethic, thrift, attitudes towards profit, trust, etc.
4. *Institutions*: Rules, regulations, laws, and policies that affect economic incentives to invest in technology, physical, and human capital. The crucial aspect is that institutions are *choices* made by society.

Let us discuss each one of them.

7.2.1 | Luck

This is essentially a catchier way of talking about multiple equilibria. If we go back to our models of poverty traps, we will recall that, in many cases, the same set of parameters is consistent with more than one equilibrium. Moreover, these equilibria can be ranked in welfare terms. As a result, it is possible (at least theoretically) that identical countries will end up in very different places.

But is the theoretical possibility that important empirically? Do we really believe that Switzerland is rich and Malawi is poor essentially because of luck? It seems a little hard to believe. Even if we go back in time, it seems that initial conditions were very different in very relevant dimensions. In other words, multiple equilibria might explain relatively small and short-lived divergence, but not the bulk of the mind-boggling cross-country differences we see – at least not in isolation.

In addition, from a conceptual standpoint, a drawback is that we need to explain the coordination failures and how people fail to coordinate even when they are trapped in a demonstrably bad equilibrium. This pushes back the explanatory challenge by another degree.

In sum, it seems that multiple equilibria and luck might be relevant, but in conjunction with other explanations. For instance, it may be that a country happened to be ruled by a growth-friendly dictator, while another was stuck with a growth-destroying one. Jones and Olken (2005) use random deaths of country leaders to show that there does seem to be an impact on subsequent performance. The question then becomes why the effects of these different rulers would matter over the long run, and for this we would have to consider some of the other classes of explanations.[7]

7.2.2 | Geography

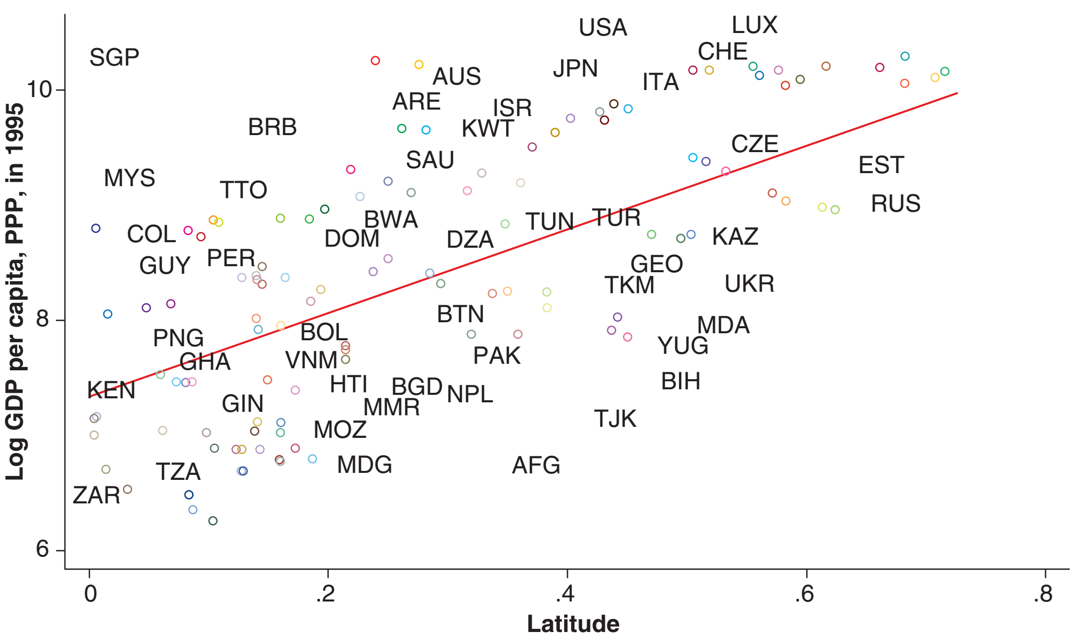
This is somewhat related to the luck hypothesis, but certainly distinctive: perhaps the deepest source of heterogeneity between countries is the natural environment they happened to be endowed with. From a very big picture perspective, geographical happenstance of this sort is a very plausible candidate for a determinant of broad development paths, as argued for instance by Jarred Diamond in his 1999 Pulitzer-Prize-winning book *Guns, Germs and Steel*[8]. As an example, Diamond suggests that one key reason Europe conquered America, and not the other way around, was that Europe had an endowment of big animal species that were relatively easy to domesticate, which in turn led to improved immunisation by humans exposed to animal-borne diseases, and more technological advances. But can geography also explain differences in economic performance at the scale on which we usually think about them, say between different countries over decades or even a couple of centuries?

On some level, it is hard to think that the natural environment would not affect economic performance, on any time frame. Whether a country is in the middle of the Sahara desert, the Amazon rain forest, or some temperate climate zone must make some difference for the set of economic opportunities that it faces. This idea becomes more compelling when we look at the correlation between certain geographical variables and economic performance, as illustrated by the Figure (7.2), again taken from Acemoglu (2009). It is clear from that picture that countries that are closer to the equator are poorer on average. At the very least, any explanation for economic performance would have to be consistent with this stylised fact. The question, once again, is to assess to what extent these geographical differences underlie the ultimate performance, and this is not an easy empirical question.

Let us start by considering the possible conceptual arguments. The earliest version of the geography hypothesis has to do with the effect of the climate on the effort – the old idea that hot climates are not conducive to hard work. While this seems very naive (and not too politically correct) to our 21st century ears, the idea that climate (and geography more broadly) affects technological productivity, especially in agriculture, still sounds very plausible. If these initial differences in turn condition subsequent technological progress (as argued by Jared Diamond, as we have seen, and as we will see, in different forms, by Jeffrey Sachs), it just might be that geography is the ultimate determinant of the divergence between societies over the very long run.

A big issue with this modern version of the geography hypothesis is that it is much more appealing to think of geography affecting agricultural productivity, but modern growth seems to have a lot more to do with industrialisation. While productivity in agriculture might have conditioned the development of industry to begin with, once industrial technologies are developed we would have to explain why they are not adopted by some countries. Geography is probably not enough to account for that, at least in this version of the story.

Figure 7.2 Distance from the equator and income, from Acemoglu (2012)



Another version has to do with the effect of geography on the disease environment, and the effect of the latter on productivity. This is a version strongly associated with Jeffrey Sachs (2002), who argues that the disease burden in the tropics (malaria in particular) can explain a lot of why Africa is so poor. The basic idea is very simple: unhealthy people are less productive. However, many of these diseases have been (or could potentially be) tamed by technological progress, so the question becomes one of why some countries have failed to benefit from that progress. In other words, the disease environment that prevails in a given country is also a consequence of its economic performance. While this doesn't mean that there cannot be causality running in the other direction, at the very least it makes the empirical assessment substantially harder.

What does the evidence say, broadly speaking? Acemoglu et al. (2002) (henceforth AJR) make the argument of the *reversal of fortune* to suggest that geography cannot explain that much. Consider the set of countries that were colonised by the Europeans, starting in the 15th century. The point is that countries that were richer before colonisation eventually became poorer – think about Peru or Mexico versus Canada, Australia, or the U.S. (see Figures 7.3 and 7.4). But geography, if the concept is to mean anything, is largely constant over time! (At least over the time periods we talk about.)

But how about the version that operates through the disease environment? This might operate on a smaller scale than the one that is belied by the reversal of fortunes argument. To assess this argument, we want to have some exogenous variation in the disease environment, that enables us to disentangle the two avenues of causality. Acemoglu and Johnson (2007) use the worldwide technological shocks that greatly improved control over many of the world's worst diseases. They measure this exogenous impact, at the country level, by considering the date at which a technological breakthrough

Figure 7.3 Reversal of fortunes - urbanization, from Acemoglu (2012)

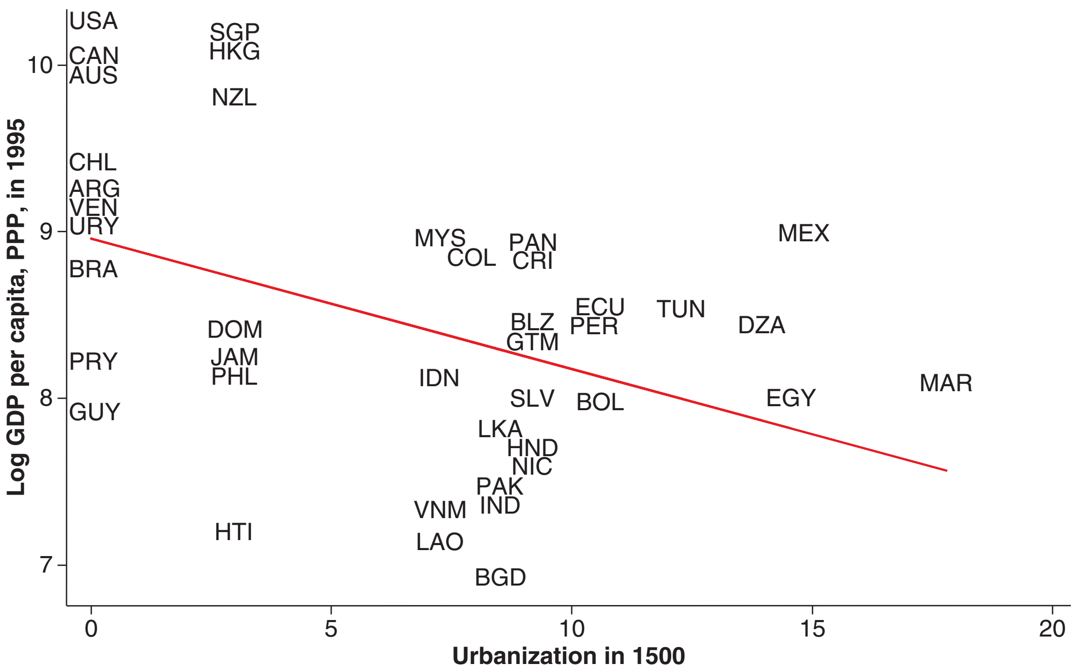
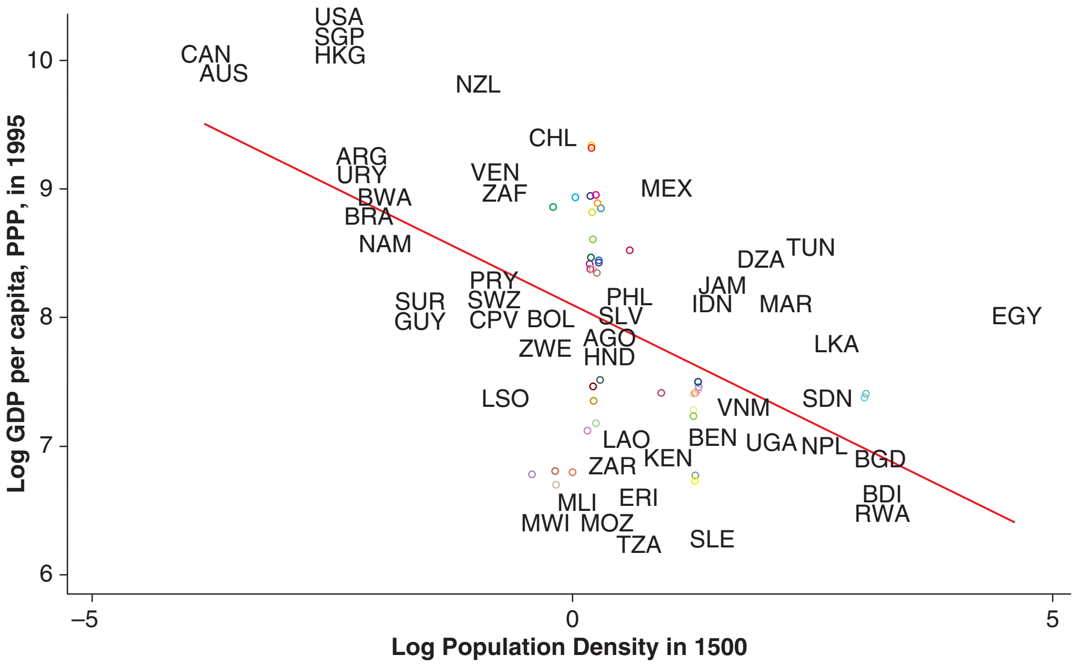


Figure 7.4 Reversal of fortunes -pop. density, from Acemoglu (2012)



was obtained against a given disease, such as tuberculosis or malaria, and the country's initial exposure to that disease. What they show is that, quite beyond not having a quantitatively important effect on output per capita, these health interventions actually seem not to have had any significant effect at all.[9]

Finally, another version of the geography argument relates to the role of natural resources and growth. Sachs and Warner (2001) tackle this issue and find a surprising result: countries endowed with natural resources seem to grow slower than countries that do not (think of Congo, Zambia or Iran, vs Japan and Hong Kong). How could this be so? Shouldn't having more resources be good? Sachs associates the poorer performance to the fact that societies that are rich in resources become societies of rent-seekers, societies where appropriating the wealth of natural resources is more important than creating new wealth. Another explanation has to do with relative prices. Commodity booms lead to a sustained real exchange rate appreciation that fosters the growth of non-tradable activities, where productivity growth seems a bit slower. Finally, commodity economies suffer large volatility in their real exchange rates, making economic activity more risky both in the tradable and non-tradable sectors. This decreases the incentives to invest, as we will see later in the book, and also hurts growth prospects. Obviously, this is not a foregone conclusion. Some countries like Norway or Chile have learnt to deal with the challenge of natural resources by setting sovereign wealth funds or investment strategies that try to diminish these negative effects. But then this, once again, pushes the question of this dimension of geography to that of institutions, to which we will shortly turn below.

7.2.3 | Culture

What do we mean by culture? The standard definition used by economists, as spelled out by Guiso et al. (2006), refers to "those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation" (p. 23). In other words, culture is something that lives inside people's heads – as opposed to being external to them – but it is not something idiosyncratic to individuals; it is built and, importantly, transmitted at the level of groups.

It is hard to argue against the assertion that people's beliefs, values, and attitudes affect their economic decisions. It is just as clear that those beliefs, values and attitudes vary across countries (and over time). From this it is easy to conclude that culture matters for economic performance, an argument that goes back at least to Max Weber's thesis that Protestant beliefs and values, emphasising hard work and thrift, and with a positive view of wealth accumulation as a signal of God's grace, were an important factor behind the development of capitalism and the modern industrial development. In his words, the "Protestant ethic" lies behind the "spirit of capitalism".

Other arguments in the same vein have suggested that certain cultural traits are more conducive to economic growth than others (David Landes is a particularly prominent proponent of this view, as in Landes (1998)), and the distribution of those traits across countries is the key variable to ultimately understand growth. "Anglo-Saxon" values are growth-promoting, compared to "Latin" or "Asian" values, and so on. More recently, Joel Mokyr (2018) has argued that Enlightenment culture was the key driving force behind the emergence of the Industrial Revolution in Europe, and hence of the so-called "Great Divergence" between that continent and the rest of the world.

A number of issues arise with such explanations. First, culture is hard to measure, and as such may lead us into the realm of tautology. A country is rich because of its favourable culture, and a favourable culture is defined as that which is held by rich countries. This doesn't get us very far in understanding the causes of good economic performance. This circularity is particularly disturbing when the same set of values (say, Confucianism) is considered inimical to growth when Asian countries perform poorly, and suddenly becomes growth-enhancing when the same countries perform well. Second, even if culture is indeed an important causal determinant of growth, we still need to figure out where it comes from if we are to consider implications for policy and predictions for future outcomes.

These empirical and conceptual challenges have now been addressed more systematically, as better data on cultural attitudes have emerged. With such data, a vibrant literature has emerged, with economists developing theories and testing their predictions on the role that specific types of values (as opposed to a generic "culture" umbrella) play in determining economic performance. Many different types of cultural attitudes have been investigated: trust, collectivism, gender roles, beliefs about fairness, etc. This literature has often exploited historical episodes – the slave trade, the formation of medieval self-governing cities, colonisation, immigration, recessions – and specific cultural practices – religious rites, civic festivities, family arrangements – to shed light on the evolution of cultural attitudes and their impact on economic outcomes. Our assessment is that this avenue of research has already borne a lot of fruit, and remains very promising for the future. (For an overview of this literature, see the surveys by Guiso et al. (2006), Alesina and Giuliano (2015), and Nunn (2020).

As an example of this research, Campante and Yanagizawa-Drott (2015) address the question of whether one specific aspect of culture, namely religious practices, affects economic growth. They do so by focusing on the specific example of Ramadan fasting (one of the pillars of Islam). To identify a causal effect of the practice, they use variation induced by the (lunar) Islamic calendar: do (exogenously) longer Ramadan fasting hours affect economic growth? The answer they find is yes, and negatively (in Muslim countries only, reassuringly). They find a substantial effect, beyond the month of Ramadan itself, which cannot be fully explained by toll exacted by the fasting, but that they attribute to changes in labour supply decisions. People also become happier, showing that there is more to life than GDP growth. These results are consistent with existing theory on the emergence of costly religious practices. They work as screening devices to prevent free riding, and the evidence shows that more religious people become more intensely engaged, while the less committed drop out. In addition, there is an effect on individual attitudes. There is a decline in levels of general trust, suggesting that religious groups may be particularly effective in generating trust. (Given that trust is associated with good economic outcomes, we may speculate about the possible long-term impact of these changes.) In short, this illustrates how we can try to find a causal effect of cultural practices on growth, as well as trying to elucidate some of the relevant mechanisms.

7.2.4 | Institutions

Last but not least, there is the view that institutions are a fundamental source of economic growth. This idea also has an old pedigree in economics, but in modern times it has been mostly associated, in its beginnings, with the work of Douglass North (who won the Nobel Prize for his work), and more recently with scholars such as Daron Acemoglu and James Robinson. From the very beginning, here is the million-dollar question: what do we mean by institutions?

North's famous characterisation is that institutions are "the rules of the game" in a society, "the humanly devised constraints that shape human interaction" (North (1990), p. 3). Here are the key elements of his argument:

- Humanly devised: Unlike geography, institutions are chosen by groups of human beings.
- Constraints: Institutions are about placing constraints on human behaviour. Once a rule is imposed, there is a cost to breaking it.
- Shape interactions: Institutions affect incentives.

OK, fair enough. But here is the *real* question: What *exactly* do we mean by institutions? A first stab at this question is to follow the Acemoglu et al. (2005) distinctions between economic and political institutions, and between *de facto* and *de jure* institutions.

The first distinction is as follows. Economic institutions are those that directly affect the economic incentives: property rights, the presence and shape of market interactions, and regulations. They are obviously important for economic growth, as they constitute the set of incentives for accumulation and technological progress. Political institutions are those that configure the process by which society makes choices: electoral systems, constitutions, the nature of political regimes, the allocation of political power etc. There is clearly an intimate connection between those two types, as political power affects the economic rules that will prevail.

The second distinction is just as important, having to do with formal vs informal rules. For instance, the law may state that all citizens have the right to vote, but in practice it might be that certain groups can have enough resources (military or otherwise) to intimidate or influence others, thereby constraining their right in practice. Formal rules, the de jure institutions, are never enough to fully characterise the rules of the game; the informal, de facto rules must be taken into consideration.

These distinctions help us structure the concepts, but we also hit the same issue that plagues the cultural explanations: since institutions are made by people, we need to understand where they come from, and how they come about. Acemoglu et al. (2005) is a great starting point to survey this literature, and (Acemoglu and Robinson 2012) provides an extremely readable overview of the ideas.

How do we assess empirically the role of institutions as a fundamental determinant of growth? At a very basic level, we can start by focusing on one thing that generates discontinuous change in institutions, but not so much in culture, and arguably not at all in geography: borders. Consider the following two examples. Figure 7.5 shows a Google Earth image of the border between Bolivia (on the left) and Brazil. We can see how the Brazilian side is fully planted with (mostly) soybeans, unlike the Bolivian side. A better-known version showing the same idea, in even starker form, is the satellite image of the Korean Peninsula at night (Figure 7.6).

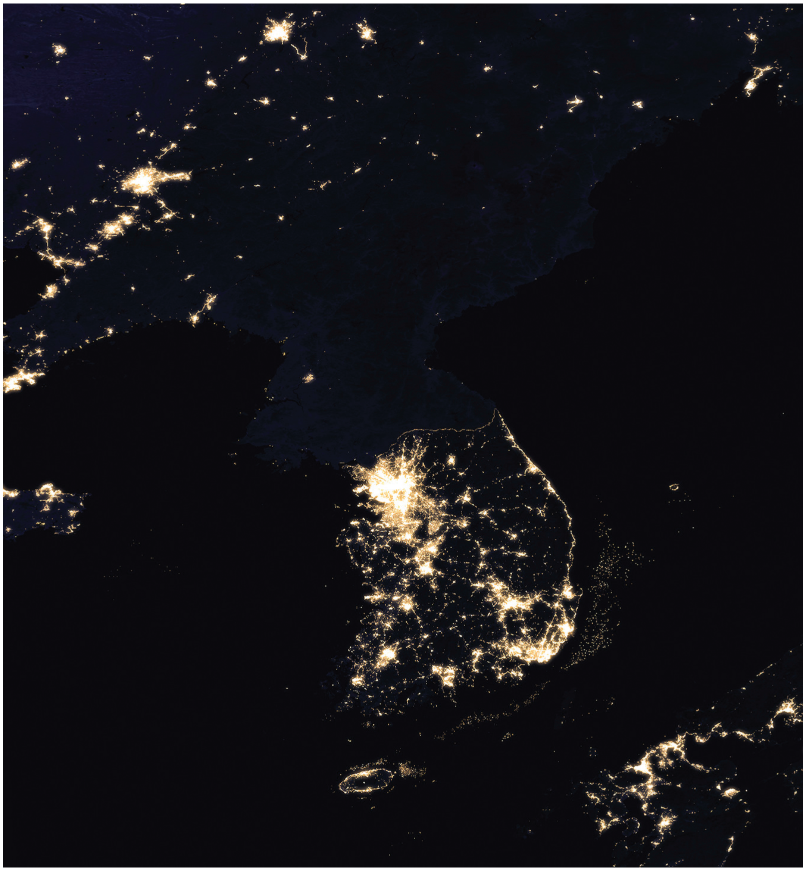
How can we do this more systematically? Here the main challenge is similar to the one facing the investigation on the effects of disease environment: is a country rich because it has good institutions, or does it have good institutions because it's rich? The seminal study here is Acemoglu et al. (2001), and it is worth going through that example in some detail – not so much for the specific answers they find, which have been vigorously debated for a couple of decades, at this point – but for how it illustrates the challenges involved, how to try and circumvent them, and the many questions that come from that process.

The paper explores the effects of a measure of institutional development given by an index of protection from expropriation. (What kind of institution is that? What are the problems with a measure like this?) The key challenge is to obtain credible exogenous variation in that measure – something that affects institutions, but not the outcome of interest (income per capita), other than through its effect on the former.

Their candidate solution for this identification problem comes again from the natural experiment of European colonisation. The argument is that current institutions are affected by the institutions that Europeans chose to implement in their colonies (persistence of institutions), and those in turn were affected by the geographical conditions they faced – in particular, the disease environment. In more inhospitable climates (from their perspective), Europeans chose not to settle, and instead set up extractive institutions. In more favourable climates they chose to settle and, as a result, ended up choosing institutions that protected the rights of the colonists. (Note that this brings in geography as a variable that affects outcomes, but *through its effect on institutions*. In particular, this helps explain the correlations with geographical variables that we observe in the data.) The key assumption is that the disease environment at the time of colonisation doesn't really affect economic outcomes today except

Figure 7.5 Border between Bolivia (left) and Brazil

Figure 7.6 The Korean Peninsula at night



through their effect on institutional development. If so, we can use variation in that environment to identify the causal effect of institutions.

Under these assumptions, they use historical measures of mortality of European settlers as an instrument for the measure of contemporaneous institutions (property rights protection), which allows them to estimate the impact of the former on contemporaneous income per capita. The resulting estimate of the impact of institutions on income per capita is 0.94. This implies that the 2.24 difference in expropriation risk between Nigeria and Chile should translate into a difference of 206 log points (approximately 8 times, since $e^{2.06} = 7.84$). So their result is that institutional factors can explain a lot of observed cross-country differences. Also, the results suggest that, once the institutional element is controlled for, there is no additional effect of the usual geographical variables, such as distance to the equator.

Their paper was extremely influential, and spawned a great deal of debate. What are some of the immediate problems with it? The most obvious is that the disease environment may have a direct impact on output (see geography hypothesis), and the disease environment back in the days of colonisation is related to that of today. They tackle this objection, and argue that the mortality of European settlers is to a large extent unrelated to the disease environment for natives, who had developed immunity to a lot of the diseases that killed those settlers. An objection that is not as obvious is whether the impact of the European settlers was through institutions, or something else. Was it culture that they brought? They argue that accounting for institutions wipes out the effect of things such as the identity of the coloniser. Was it human capital? Glaeser et al. (2004) argue that what they brought was not good institutions, but themselves: the key was their higher levels of human capital, which in turn are what is behind good institutions. This is a much harder claim to settle empirically, so the question remains open.

Broadly speaking, there is wide acceptance in the profession, these days, that institutions play an important role in economic outcomes. However, there is a lot of room for debate as to which are the relevant institutions, and where they come from. How do societies choose different sets of institutions? Particularly if some institutions are better for economic performance than others, why do some countries choose bad institutions? Could it be because some groups benefit from inefficient institutions? If so, how do they manage to impose them on the rest of society? In other words, we need to understand the political economy of institutional development. This is a very open and exciting area of research, both theoretically and empirically.

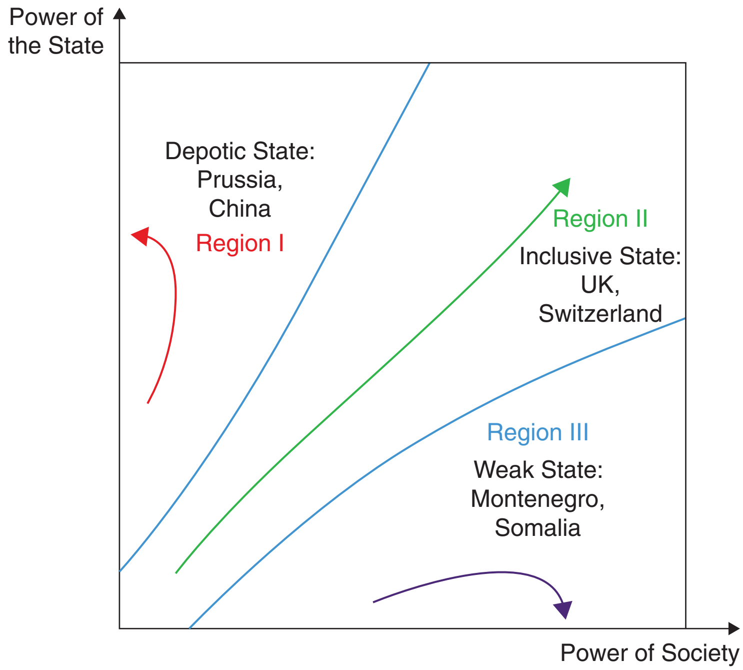
As an example of the themes in the literature, Acemoglu and Robinson (2019) asks not only why certain countries develop a highly capable state and others don't, but also why, among those that do, some have that same state guarantee the protection of individual rights and liberties, while others have a state that tramples on those rights and liberties. Their argument is that institutional development proceeds as a race between the power of the state and the power of society, as people both demand the presence of the Leviathan enforcing rules and order, and resent its power. If the state gets too powerful relative to society, the result is a despotic situation; if the state is too weak, the result is a state incapable of providing the needed underpinnings for development. In the middle, there is the "narrow corridor" along which increasing state capacity pushes for more societal control, and the increased power of society pushes for a more capable (and inclusive) state. The dynamics are illustrated by Figure 7.7, and one crucial aspect is worth mentioning: small differences in initial conditions – say, two economies just on opposite sides of the border between two regions in the figure – can evolve into vastly different institutional and development paths.

7.3 | What have we learned?

When it comes to the proximate causes of growth, in spite of the limitations of each specific empirical approach – growth accounting, regression methods, and calibration – the message from the data is reasonably clear, yet nuanced: factor accumulation can arguably explain a substantial amount of income differences, and specific growth episodes, but ultimately differences in productivity are very important. This is a bit daunting, since the fact is that we don't really understand what productivity *is*, in a deeper sense. Still, it underscores the importance of the process of technological progress – and the policy issues raised in Chapter 6 – as a primary locus for growth policies.

How about the fundamental causes? There is certainly a role for geography and luck (multiple equilibria), but our reading of the literature is that culture and institutions play a key part. There remains a lot to be learned about how these things evolve, and how they affect outcomes, and these are bound to be active areas of research for the foreseeable future.

Figure 7.7 Weak, despotic and inclusive states, from Acemoglu and Robinson (2017)



7.4 | What next?

Once again, the growth textbook by Acemoglu (2009) is a superb resource, and it contains a more in-depth discussion of the empirical literature on the proximate causes of growth. It also has a very interesting discussion on the fundamental causes, but it's useful to keep in mind that, its author being one of the leading proponents of the view that institutions matter most, it certainly comes at that debate from that specific point of view.

Specifically on culture, the best places to go next are the survey articles we mentioned in our discussion. The survey by Guiso et al. (2006) is a bit outdated, of course, but still a great starting point. The more recent surveys by Alesina and Giuliano (2015), focusing particularly on the links between culture and institutions, and by Nunn (2020), focusing on the work using historical data, are very good guides to where the literature is and is going.

On institutions, there is no better place to go next than the books by Acemoglu and Robinson (2012) and Acemoglu and Robinson (2019). They are very ambitious intellectual exercises, encompassing theory, history, and empirical evidence, and meant for a broad audience – which makes them a fun and engaging read.

These being very active research fields, there are a lot of questions that remain open. Anyone interested in the social sciences, as the readers of this book most likely are, will find a lot of food for thought in these sources.

Notes

[1] We know, of course, that the NGM does not generate long-run growth, except through exogenous technical progress. However, keep in mind that we also live in the transition path!

[2] Measuring each of these variables is an art by itself, and hundreds of papers have tried to refine these measures. Capital stocks are usually computed from accumulating past net investment and human capital from accumulating population adjusted by their productivity, assessed through Mincer equations relating years of schooling and income.

[3] This memorable phrase is attributed to Moses Abramovitz.

[4] Check out the priceless first paragraph of his 1995 paper summarising his findings: "This is a fairly boring and tedious paper, and is intentionally so. This paper provides no new interpretations of the East Asian experience to interest the historian, derives no new theoretical implications of the forces behind the East Asian growth process to motivate the theorist, and draws no new policy implications from the subtleties of East Asian government intervention to excite the policy activist. Instead, this paper concentrates its energies on providing a careful analysis of the historical patterns of output growth, factor accumulation, and productivity growth in the newly industrializing countries (NICs) of East Asia, i.e., Hong Kong, Singapore, South Korea, and Taiwan" (p. 640).

[5] As we mentioned, Kremer et al. (2021) have argued that the data has moved in the direction of absolute convergence across countries in the 21st century.

[6] This allegiance is also behind their just as playful title, "A Contribution to the Empirics of Economic Growth", which substitutes empirics for the theory from Solow's original article.

[7] For instance, the aforementioned work by Jones and Olken (2005) shows that the effect of leaders is present in non-democracies, but not in democracies, suggesting that luck of this sort may matter insofar as it interacts with (in this case) institutional features.

[8] Diamond (2013).

[9] How can that be? Think about what happens, in the context of the Solow model, when population increases.

References

Acemoglu, D. (2009). *Introduction to modern economic growth*. Princeton University Press.

Acemoglu, D. & Dell, M. (2010). Productivity differences between and within countries. *American Economic Journal: Macroeconomics*, *2*(1), 169–88.

Acemoglu, D. & Johnson, S. (2007). Disease and development: The effect of life expectancy on economic growth. *Journal of Political Economy*, *115*(6), 925–985.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91*(5), 1369–1401.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *The Quarterly Journal of Economics*, *117*(4), 1231–1294.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. *Handbook of Economic Growth*, *1*, 385–472.

Acemoglu, D. & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Currency.

Acemoglu, D. & Robinson, J. A. (2017). *The emergence of weak, despotic and inclusive states*. National Bureau of Economic Research.

Acemoglu, D. & Robinson, J. A. (2019). *The narrow corridor: How nations struggle for liberty*. Penguin UK.

Alesina, A. & Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, *53*(4), 898–944.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics*, *106*(2), 407–443.

Barro, R. J. & Sala-i-Martin, X. (1992). Convergence. *Journal of Political Economy*, *100*(2), 223–251.

Baumol, W. J. (1986). Productivity growth, convergence, and welfare: What the long-run data show. *The American Economic Review*, 1072–1085.

Campante, F. & Yanagizawa-Drott, D. (2015). Does religion affect economic growth and happiness? Evidence from Ramadan. *The Quarterly Journal of Economics*, *130*(2), 615–658.

De Long, J. B. (1988). Productivity growth, convergence, and welfare: Comment. *The American Economic Review*, *78*(5), 1138–1154.

Diamond, J. (2013). *Guns, germs and steel: A short history of everybody for the last 13,000 years*. Random House.

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, *9*(3), 271–303.

Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, *20*(2), 23–48.

Hall, R. E. & Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, *114*(1), 83–116.

Jones, B. F. & Olken, B. A. (2005). Do leaders matter? National leadership and growth since World War II. *The Quarterly Journal of Economics*, *120*(3), 835–864.

Kremer, M., Willis, J., & You, Y. (2021). Converging to convergence. *NBER Macro Annual 2021*. https://www.nber.org/system/files/chapters/c14560/c14560.pdf.

Krugman, P. (1994). The myth of Asia's miracle. *Foreign Affairs*, *73*(6), 62–78.

Landes, D. S. (1998). Culture counts. *Challenge*, *41*(4), 14–30.

Levine, R. & Renelt, D. (1991). *Cross-country studies of growth and policy: Methodological, conceptual, and statistical problems* (Vol. 608). World Bank Publications.

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics*, *107*(2), 407–437.

Mokyr, J. (2018). *The British industrial revolution: An economic perspective*. Routledge.

North, D. (1990). Institutions, institutional change and economic performance. Cambridge University Press: New York.

Nunn, N. (2020). The historical roots of economic development. *Science*, *367*(6485).

Rodrik, D. (2012). Why we learn nothing from regressing economic growth on policies. *Seoul Journal of Economics*, *25*.

Sachs, J. & Malaney, P. (2002). The economic and social burden of malaria. *Nature*, *415*(6872), 680–685.

Sachs, J. D. & Warner, A. M. (2001). The curse of natural resources. *European Economic Review*, *45*(4-6), 827–838.

Sala-i-Martin, X. X. (1997). *I just ran four million regressions* (tech. rep.) National Bureau of Economic Research.

Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, 312–320.

Summers, R. & Heston, A. (1988). A new set of international comparisons of real product and price levels estimates for 130 countries, 1950–1985. *Review of Income and Wealth*, *34*(1), 1–25.

Young, A. (1995). The tyranny of numbers: Confronting the statistical realities of the east asian growth experience. *The Quarterly Journal of Economics*, *110*(3), 641–680.