1

Economic Growth and Economic Development: The Questions

1.1 Cross-Country Income Differences

There are very large differences in income per capita and output per worker across countries today. Countries at the top of the world income distribution are more than 30 times as rich as those at the bottom. For example, in 2000, gross domestic product (GDP; or income) per capita in the United States was more than $34,000. In contrast, income per capita is much lower in many other countries: about $8,000 in Mexico, about $4,000 in China, just over $2,500 in India, only about $1,000 in Nigeria, and much, much lower in some other sub-Saharan African countries, such as Chad, Ethiopia, and Mali. These numbers are all in 2000 U.S. dollars and are adjusted for purchasing power parity (PPP) to allow for differences in relative prices of different goods across countries.[1] The cross-country income gap is considerably larger when there is no PPP adjustment. For example, without the PPP adjustment, GDP per capita in India and China relative to the United States in 2000 would be lower by a factor of four or so.

Figure 1.1 provides a first look at these differences. It plots estimates of the distribution of PPP-adjusted GDP per capita across the available set of countries in 1960, 1980, and 2000. A number of features are worth noting. First, the 1960 density shows that 15 years after the end of World War II, most countries had income per capita less than $1,500 (in 2000 U.S. dollars); the mode of the distribution is around $1,250. The rightward shift of the distributions for 1980 and 2000 shows the growth of average income per capita for the next 40 years. In 2000, the mode is slightly above $3,000, but now there is another concentration of countries between $20,000 and $30,000. The density estimate for the year 2000 shows the considerable inequality in income per capita today.

The spreading out of the distribution in Figure 1.1 is partly because of the increase in average incomes. It may therefore be more informative to look at the logarithm (log) of

1. All data are from the Penn World tables compiled by Heston, Summers, and Aten (2002). Details of data sources and more on PPP adjustment can be found in the References and Literature section at the end of this chapter.

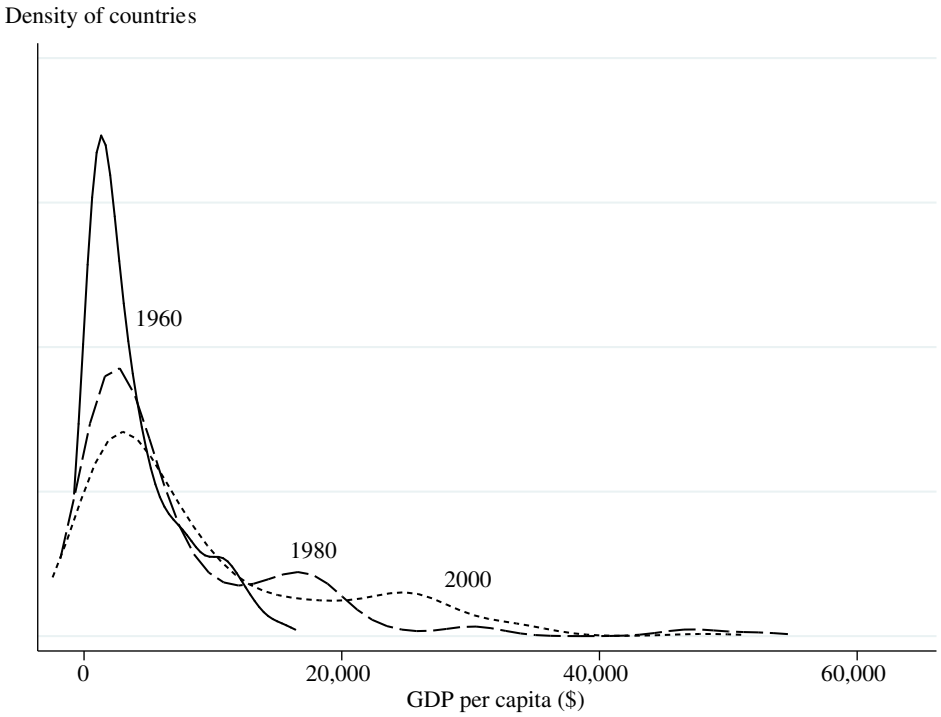

FIGURE 1.1 Estimates of the distribution of countries according to PPP-adjusted GDP per capita in 1960, 1980, and 2000.

income per capita. It is more natural to look at the log of variables, such as income per capita, that grow over time, especially when growth is approximately proportional, as suggested by Figure 1.8 below. This is for the simple reason that when $x\,(t)$ grows at a proportional rate, $\log x\,(t)$ grows linearly, and if $x_1\,(t)$ and $x_2\,(t)$ both grow by the same proportional amount, $\log x_1\,(t) - \log x_2\,(t)$ remains constant, while $x_1\,(t) - x_2\,(t)$ increases.

Figure 1.2 shows a similar pattern, but now the spreading is more limited, because the absolute gap between rich and poor countries has increased considerably between 1960 and 2000, while the proportional gap has increased much less. Nevertheless, it can be seen that the 2000 density for log GDP per capita is still more spread out than the 1960 density. In particular, both figures show that there has been a considerable increase in the density of relatively rich countries, while many countries still remain quite poor. This last pattern is sometimes referred to as the "stratification phenomenon," corresponding to the fact that some of the middle-income countries of the 1960s have joined the ranks of relatively high-income countries, while others have maintained their middle-income status or even experienced relative impoverishment.

Figures 1.1 and 1.2 demonstrate that there is somewhat greater inequality among nations today than in 1960. An equally relevant concept might be inequality among individuals in the world economy. Figures 1.1 and 1.2 are not directly informative on this, since they treat each country identically regardless of the size of its population. An alternative is presented in Figure 1.3, which shows the population-weighted distribution. In this case, countries such as China, India, the United States, and Russia receive greater weight because they have larger populations. The picture that emerges in this case is quite different. In fact, the 2000 distribution looks less spread out, with a thinner left tail than the 1960 distribution. This reflects the fact that

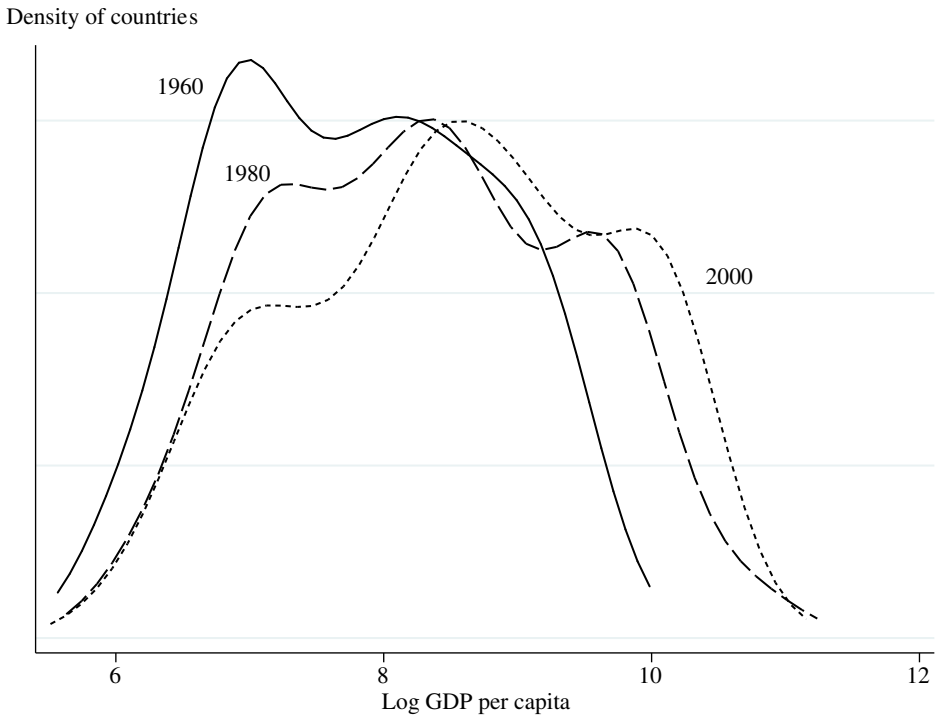

FIGURE 1.2 Estimates of the distribution of countries according to log GDP per capita (PPP adjusted) in 1960, 1980, and 2000.

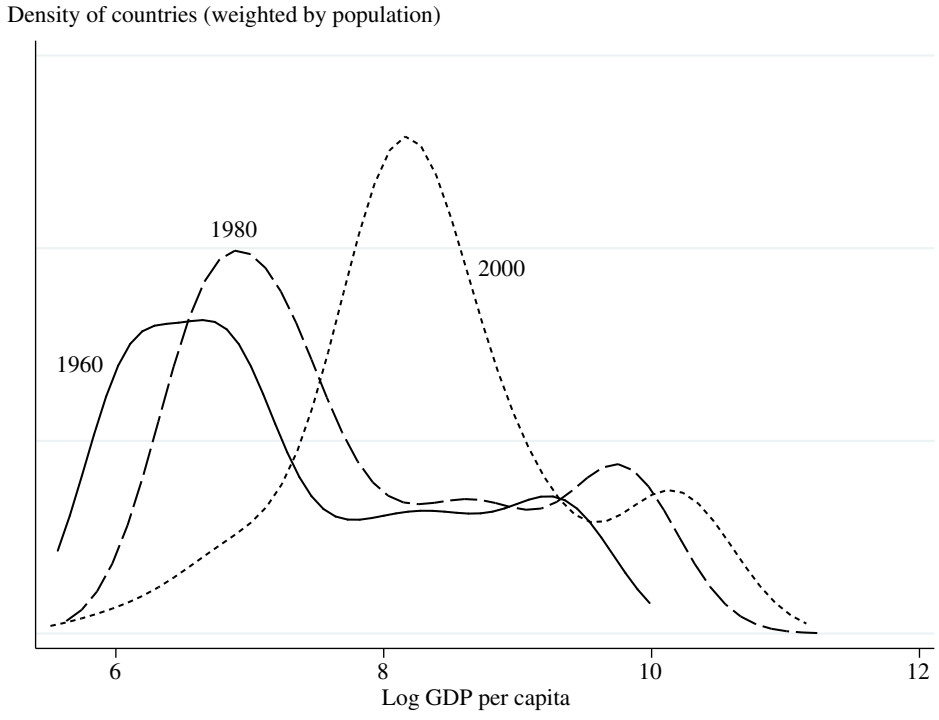

FIGURE 1.3 Estimates of the population-weighted distribution of countries according to log GDP per capita (PPP adjusted) in 1960, 1980, and 2000.

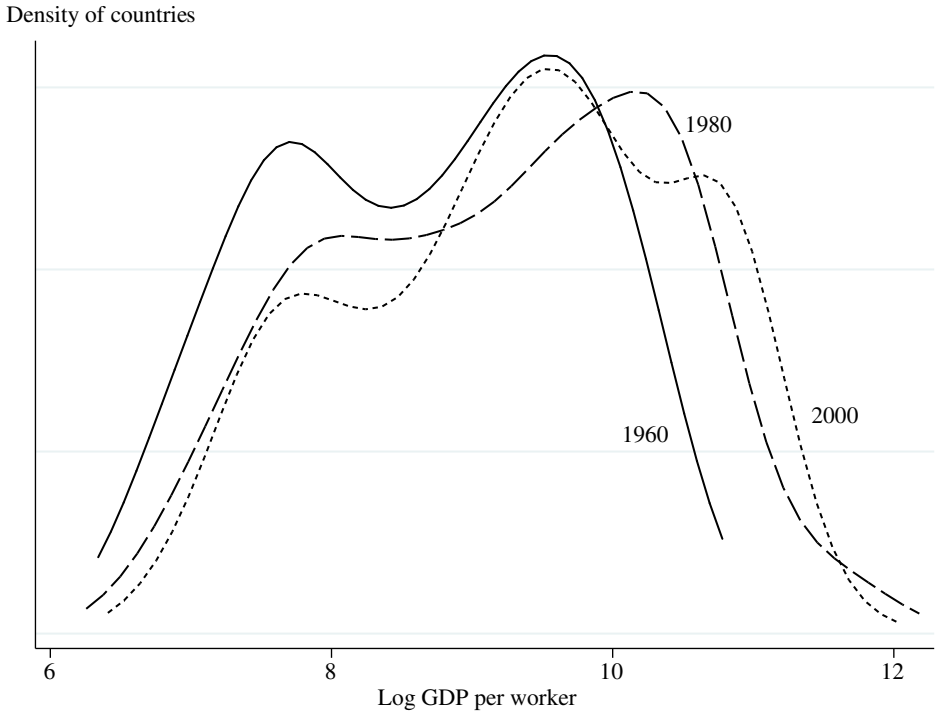

FIGURE 1.4 Estimates of the distribution of countries according to log GDP per worker (PPP adjusted) in 1960, 1980, and 2000.

in 1960 China and India were among the poorest nations in the world, whereas their relatively rapid growth in the 1990s puts them into the middle-poor category by 2000. Chinese and Indian growth has therefore created a powerful force for relative equalization of income per capita among the inhabitants of the globe.

Figures 1.1, 1.2, and 1.3 look at the distribution of GDP per capita. While this measure is relevant for the welfare of the population, much of growth theory focuses on the productive capacity of countries. Theory is therefore easier to map to data when we look at output (GDP) per worker. Moreover, key sources of difference in economic performance across countries are national policies and institutions. So for the purpose of understanding the sources of differences in income and growth across countries (as opposed to assessing welfare questions), the unweighted distribution is more relevant than the population-weighted distribution. Consequently, Figure 1.4 looks at the unweighted distribution of countries according to (PPP-adjusted) GDP per worker. "Workers" here refers to the total economically active population (according to the definition of the International Labour Organization). Figure 1.4 is very similar to Figure 1.2, and if anything, it shows a greater concentration of countries in the relatively rich tail by 2000, with the poor tail remaining more or less the same as in Figure 1.2.

Overall, Figures 1.1–1.4 document two important facts: first, there is great inequality in income per capita and income per worker across countries as shown by the highly dispersed distributions. Second, there is a slight but noticeable increase in inequality across nations (though not necessarily across individuals in the world economy).

1.2 Income and Welfare

Should we care about cross-country income differences? The answer is definitely yes. High income levels reflect high standards of living. Economic growth sometimes increases pollution or may raise individual aspirations, so that the same bundle of consumption may no longer satisfy an individual. But at the end of the day, when one compares an advanced, rich country with a less-developed one, there are striking differences in the quality of life, standards of living, and health.

Figures 1.5 and 1.6 give a glimpse of these differences and depict the relationship between income per capita in 2000 and consumption per capita and life expectancy at birth in the same year. Consumption data also come from the Penn World tables, while data on life expectancy at birth are available from the World Bank Development Indicators.

These figures document that income per capita differences are strongly associated with differences in consumption and in health as measured by life expectancy. Recall also that these numbers refer to PPP-adjusted quantities; thus differences in consumption do not (at least in principle) reflect the differences in costs for the same bundle of consumption goods in different countries. The PPP adjustment corrects for these differences and attempts to measure the variation in real consumption. Thus the richest countries are not only producing more than 30 times as much as the poorest countries, but are also consuming 30 times as much. Similarly, cross-country differences in health are quite remarkable; while life expectancy at birth is as

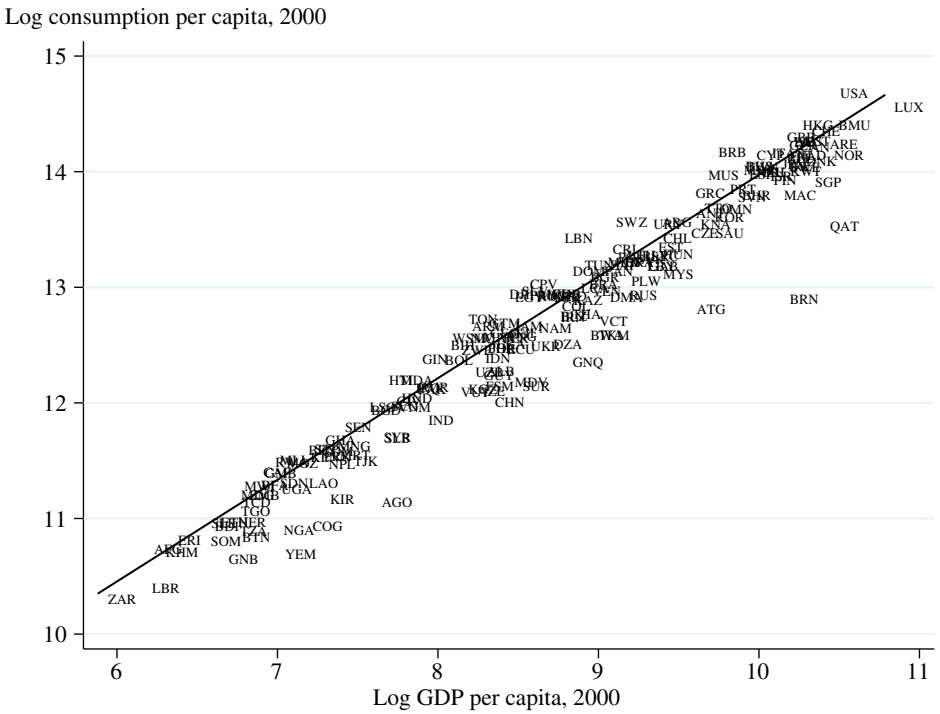

FIGURE 1.5 The association between income per capita and consumption per capita in 2000. For a definition of the abbreviations used in this and similar figures in the book, see http://unstats.un.org/unsd/methods/m49/m49alpha.htm.

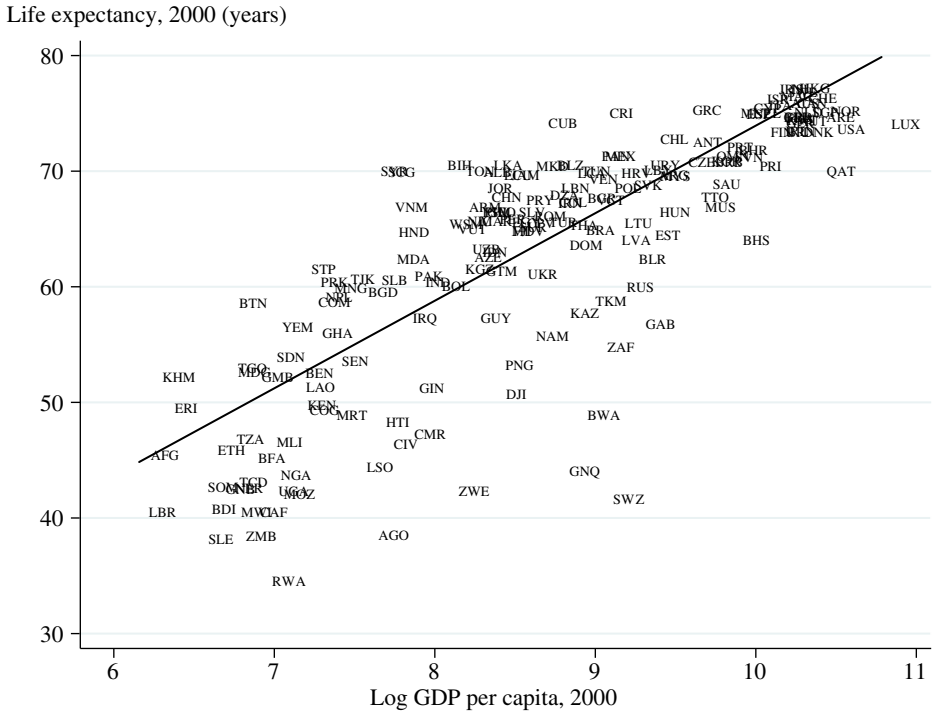

FIGURE 1.6 The association between income per capita and life expectancy at birth in 2000.

high as 80 in the richest countries, it is only between 40 and 50 in many sub-Saharan African nations. These gaps represent huge welfare differences.

Understanding why some countries are so rich while some others are so poor is one of the most important, perhaps *the* most important, challenges facing social science. It is important both because these income differences have major welfare consequences and because a study of these striking differences will shed light on how the economies of different nations function and how they sometimes fail to function.

The emphasis on income differences across countries implies neither that income per capita can be used as a "sufficient statistic" for the welfare of the average citizen nor that it is the only feature that we should care about. As discussed in detail later, the efficiency properties of the market economy (such as the celebrated First Welfare Theorem or Adam Smith's invisible hand) do not imply that there is no conflict among individuals or groups in society. Economic growth is generally good for welfare but it often creates winners and losers. Joseph Schumpeter's famous notion of creative destruction emphasizes precisely this aspect of economic growth; productive relationships, firms, and sometimes individual livelihoods will be destroyed by the process of economic growth, because growth is brought about by the introduction of new technologies and creation of new firms, replacing existing firms and technologies. This process creates a natural social tension, even in a growing society. Another source of social tension related to growth (and development) is that, as emphasized by Simon Kuznets and discussed in detail in Part VII, growth and development are often accompanied by sweeping structural transformations, which can also destroy certain established relationships and create yet other winners and losers in the process. One of the important questions of

political economy, which is discussed in the last part of the book, concerns how institutions and policies can be arranged so that those who lose out from the process of economic growth can be compensated or prevented from blocking economic progress via other means.

A stark illustration of the fact that growth does not always mean an improvement in the living standards of all or even most citizens in a society comes from South Africa under apartheid. Available data (from gold mining wages) suggest that from the beginning of the twentieth century until the fall of the apartheid regime, GDP per capita grew considerably, but the real wages of black South Africans, who make up the majority of the population, likely fell during this period. This of course does not imply that economic growth in South Africa was not beneficial. South Africa is still one of the richest countries in sub-Saharan Africa. Nevertheless, this observation alerts us to other aspects of the economy and also underlines the potential conflicts inherent in the growth process. Similarly, most existing evidence suggests that during the early phases of the British industrial revolution, which started the process of modern economic growth, the living standards of the majority of the workers may have fallen or at best remained stagnant. This pattern of potential divergence between GDP per capita and the economic fortunes of large numbers of individuals and society is not only interesting in and of itself, but it may also inform us about why certain segments of the society may be in favor of policies and institutions that do not encourage growth.

1.3 Economic Growth and Income Differences

How can one country be more than 30 times richer than another? The answer lies in differences in growth rates. Take two countries, A and B, with the same initial level of income at some date. Imagine that country A has 0% growth per capita, so its income per capita remains constant, while country B grows at 2% per capita. In 200 years' time country B will be more than 52 times richer than country A. This calculation suggests that the United States might be considerably richer than Nigeria because it has grown steadily over an extended period of time, while Nigeria has not. We will see that there is a lot of truth to this simple calculation. In fact, even in the historically brief postwar era, there are tremendous differences in growth rates across countries. These differences are shown in Figure 1.7 for the postwar era, which plots the density of growth rates across countries in 1960, 1980, and 2000. The growth rate in 1960 refers to the (geometric) average of the growth rate between 1950 and 1969, the growth rate in 1980 refers to the average growth rate between 1970 and 1989, and 2000 refers to the average between 1990 and 2000 (in all cases subject to data availability). Figure 1.7 shows that in each time interval, there is considerable variability in growth rates; the cross-country distribution stretches from negative rates to average rates as high as 10% per year. It also shows that average growth in the world was more rapid in the 1950s and 1960s than in the subsequent decades.

Figure 1.8 provides another look at these patterns by plotting log GDP per capita for a number of countries between 1960 and 2000 (in this case, I plot GDP per capita instead of GDP per worker because of the availability of data and to make the figures more comparable to the historical figures below). At the top of the figure, U.S. and U.K. GDP per capita increase at a steady pace, with a slightly faster growth in the United States, so that the log (or proportional) gap between the two countries is larger in 2000 than it is in 1960. Spain starts much poorer than the United States and the United Kingdom in 1960 but grows very rapidly between 1960 and the mid-1970s, thus closing the gap between itself and the latter two countries. The three countries that show the most rapid growth in this figure are Singapore, South Korea, and Botswana. Singapore starts much poorer than the United Kingdom and Spain in 1960 but

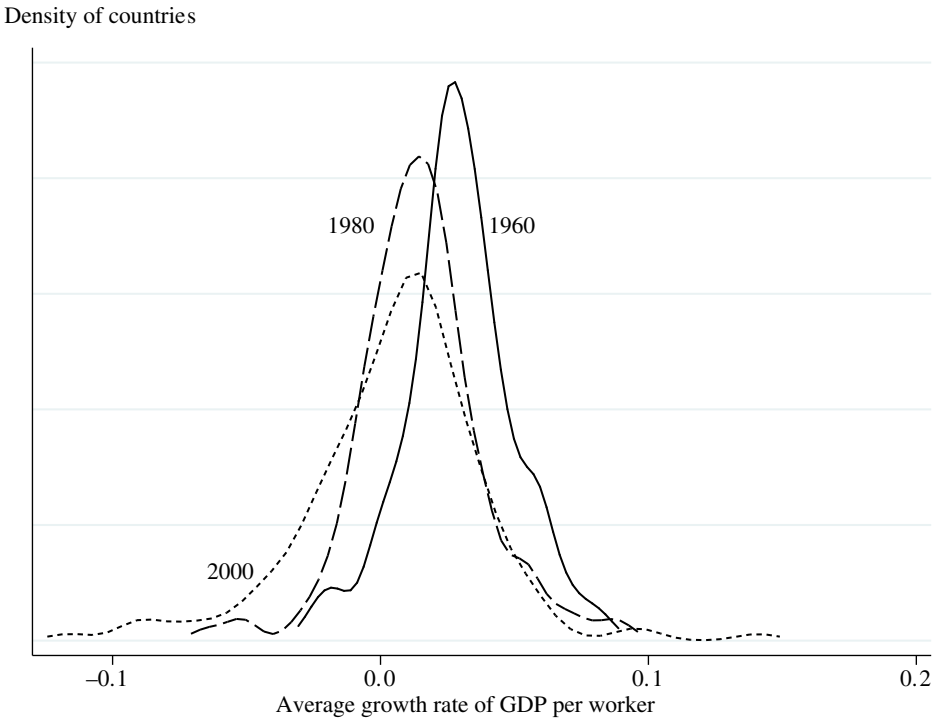

FIGURE 1.7 Estimates of the distribution of countries according to the growth rate of GDP per worker (PPP adjusted) in 1960, 1980, and 2000.

grows rapidly, and by the mid-1990s it has become richer than both. South Korea has a similar trajectory, though it starts out poorer than Singapore and grows slightly less rapidly, so that by the end of the sample it is still a little poorer than Spain. The other country that has grown very rapidly is the "African success story" Botswana, which was extremely poor at the beginning of the sample. Its rapid growth, especially after 1970, has taken Botswana to the ranks of the middle-income countries by 2000.

The two Latin American countries in this picture, Brazil and Guatemala, illustrate the often-discussed Latin American economic malaise of the postwar era. Brazil starts out richer than South Korea and Botswana and has a relatively rapid growth rate between 1960 and 1980. But it experiences stagnation from 1980 on, so that by the end of the sample South Korea and Botswana have become richer than Brazil. Guatemala's experience is similar but even more bleak. Contrary to Brazil, there is little growth in Guatemala between 1960 and 1980 and no growth between 1980 and 2000.

Finally, Nigeria and India start out at similar levels of income per capita as Botswana but experience little growth until the 1980s. Starting in 1980, the Indian economy experiences relatively rapid growth, though this has not been sufficient for its income per capita to catch up with the other nations in the figure. Finally, Nigeria, in a pattern that is unfortunately all too familiar in sub-Saharan Africa, experiences a contraction of its GDP per capita, so that in 2000 it is in fact poorer than it was in 1960.

The patterns shown in Figure 1.8 are what we would like to understand and explain. Why is the United States richer in 1960 than other nations and able to grow at a steady pace thereafter? How did Singapore, South Korea, and Botswana manage to grow at a relatively rapid pace for

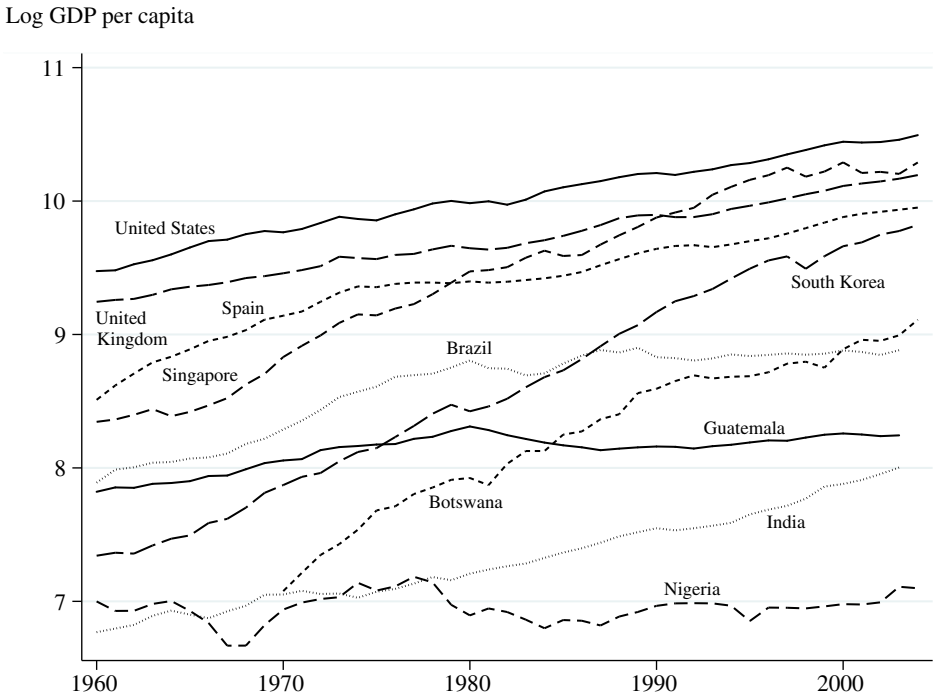

FIGURE 1.8 The evolution of income per capita in the United States, the United Kingdom, Spain, Singapore, Brazil, Guatemala, South Korea, Botswana, Nigeria, and India, 1960–2000.

40 years? Why did Spain grow relatively rapidly for about 20 years but then slow down? Why did Brazil and Guatemala stagnate during the 1980s? What is responsible for the disastrous growth performance of Nigeria?

1.4 Origins of Today's Income Differences and World Economic Growth

The growth rate differences shown in Figures 1.7 and 1.8 are interesting in their own right and could also be, in principle, responsible for the large differences in income per capita we observe today. But are they? The answer is largely no. Figure 1.8 shows that in 1960 there was already a very large gap between the United States on the one hand and India and Nigeria on the other.

This pattern can be seen more easily in Figure 1.9, which plots log GDP per worker in 2000 versus log GDP per capita in 1960 (in both cases relative to the U.S. value) superimposed over the 45° line. Most observations are around the 45° line, indicating that the relative ranking of countries has changed little between 1960 and 2000. Thus the origins of the very large income differences across nations are not to be found in the postwar era. There are striking growth differences during the postwar era, but the evidence presented so far suggests that world income distribution has been more or less stable, with a slight tendency toward becoming more unequal.

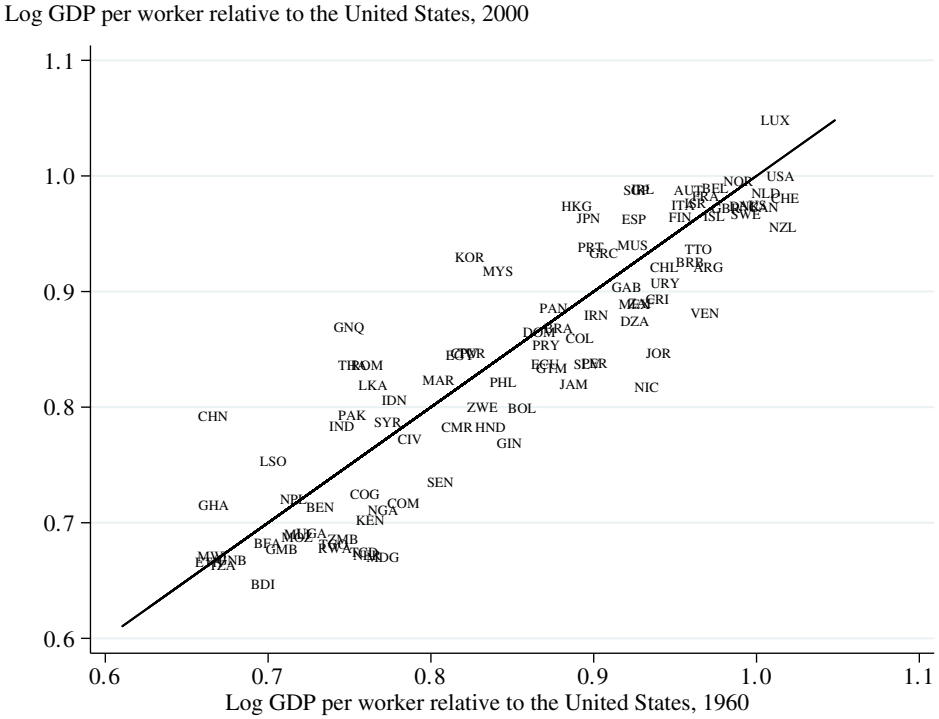

FIGURE 1.9 Log GDP per worker in 2000 versus log GDP per worker in 1960, together with the 45° line.

If not in the postwar era, when did this growth gap emerge? The answer is that much of the divergence took place during the nineteenth and early twentieth centuries. Figures 1.10–1.12 give a glimpse of these developments by using the data compiled by Angus Maddison for GDP per capita differences across nations going back to 1820 (or sometimes earlier). These data are less reliable than Summers-Heston's Penn World tables, since they do not come from standardized national accounts. Moreover, the sample is more limited and does not include observations for all countries going back to 1820. Finally, while these data include a correction for PPP, this is less complete than the price comparisons used to construct the price indices in the Penn World tables. Nevertheless, these are the best available estimates for differences in prosperity across a large number of nations beginning in the nineteenth century.

Figure 1.10 illustrates the divergence. It depicts the evolution of average income among five groups of countries: Africa, Asia, Latin America, Western Europe, and Western offshoots of Europe (Australia, Canada, New Zealand, the United States). It shows the relatively rapid growth of the Western offshoots and West European countries during the nineteenth century, while Asia and Africa remained stagnant and Latin America showed little growth. The relatively small (proportional) income gap in 1820 had become much larger by 1960.

Another major macroeconomic fact is visible in Figure 1.10: Western offshoots and West European nations experience a noticeable dip in GDP per capita around 1929 because of the famous Great Depression. Western offshoots, in particular the United States, only recovered fully from this large recession in the wake of World War II. How an economy can experience a sharp decline in output and how it recovers from such a shock are among the major questions of macroeconomics.

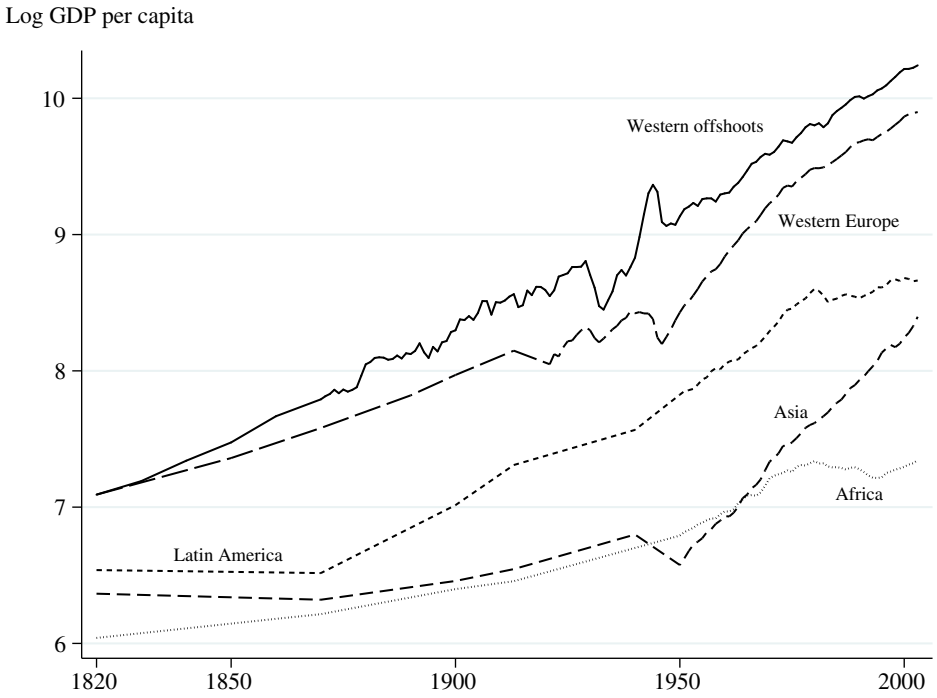

FIGURE 1.10 The evolution of average GDP per capita in Western offshoots, Western Europe, Latin America, Asia, and Africa, 1820–2000.

A variety of evidence suggests that differences in income per capita were even smaller before 1820. Maddison also has estimates for average income for the same groups of countries going back to 1000 A.D. or even earlier. Figure 1.10 can be extended back in time using these data; the results are shown in Figure 1.11. Although these numbers are based on scattered evidence and informed guesses, the general pattern is consistent with qualitative historical evidence and the fact that income per capita in any country cannot have been much less than $500 in terms of 2000 U.S. dollars, since individuals could not survive with real incomes much less than this level. Figure 1.11 shows that as we go further back in time, the gap among countries becomes much smaller. This further emphasizes that the big divergence among countries has taken place over the past 200 years or so. Another noteworthy feature that becomes apparent from this figure is the remarkable nature of world economic growth. Much evidence suggests that there was only limited economic growth before the eighteenth century and certainly before the fifteenth century. While certain civilizations, including ancient Greece, Rome, China, and Venice, managed to grow, their growth was either not sustained (thus ending with collapses and crises) or progressed only at a slow pace. No society before nineteenth-century Western Europe and the United States achieved steady growth at comparable rates.

Notice also that Maddison's estimates show a slow but steady increase in West European GDP per capita even earlier, starting in 1000. This assessment is not shared by all economic historians, many of whom estimate that there was little increase in income per capita before 1500 or even before 1800. For our purposes this disagreement is not central, however. What is important is that, using Walter Rostow's terminology, Figure 1.11 shows a pattern of *takeoff* into sustained growth; the economic growth experience of Western Europe and Western offshoots appears to have changed dramatically about 200 years or so ago. Economic historians also debate whether there was a discontinuous change in economic activity that deserves the

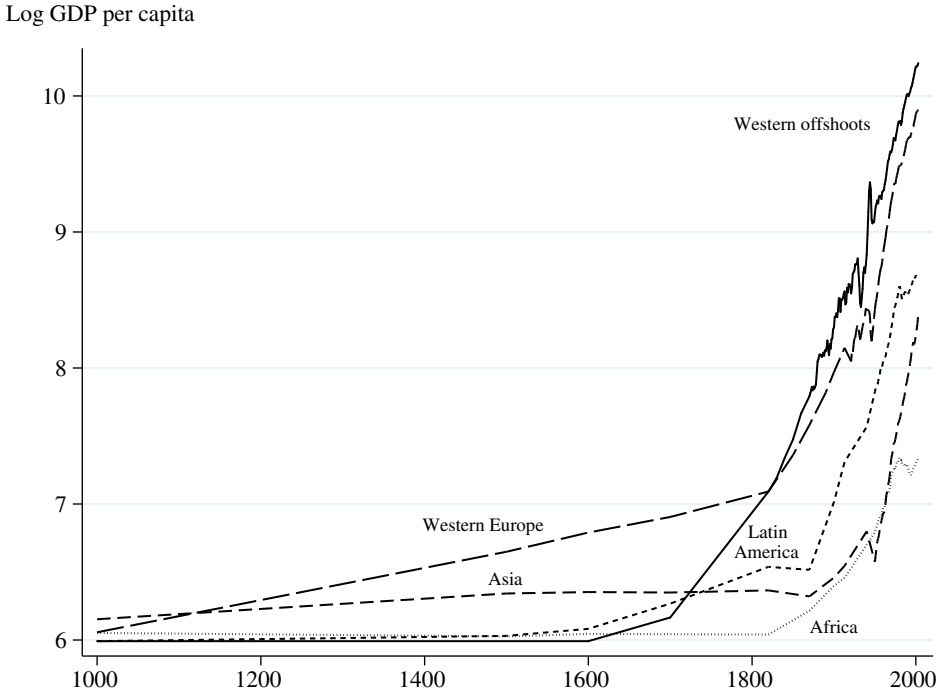

FIGURE 1.11 The evolution of average GDP per capita in Western offshoots, Western Europe, Latin America, Asia, and Africa, 1000–2000.

terms "takeoff" or "industrial revolution." This debate is again secondary to our purposes. Whether or not the change was discontinuous, it was present and transformed the functioning of many economies. As a result of this transformation, the stagnant or slowly growing economies of Europe embarked upon a path of sustained growth. The origins of today's riches and also of today's differences in prosperity are to be found in this pattern of takeoff during the nineteenth century. In the same time that Western Europe and its offshoots grew rapidly, much of the rest of the world did not experience a comparable takeoff (or did so much later). Therefore an understanding of modern economic growth and current cross-country income differences ultimately necessitates an inquiry into the causes of why the takeoff occurred, why it did so about 200 years ago, and why it took place only in some areas and not in others.

Figure 1.12 shows the evolution of income per capita for the United States, the United Kingdom, Spain, Brazil, China, India, and Ghana. This figure confirms the patterns shown in Figure 1.10 for averages, with the United States, the United Kingdom, and Spain growing much faster than India and Ghana throughout, and also much faster than Brazil and China except during the growth spurts experienced by these two countries.

Overall, on the basis of the available information we can conclude that the origins of the current cross-country differences in income per capita are in the nineteenth and early twentieth centuries (or perhaps even during the late eighteenth century). This cross-country divergence took place at the same time as a number of countries in the world "took off" and achieved sustained economic growth. Therefore understanding the origins of modern economic growth are not only interesting and important in their own right, but also holds the key to understanding the causes of cross-country differences in income per capita today.

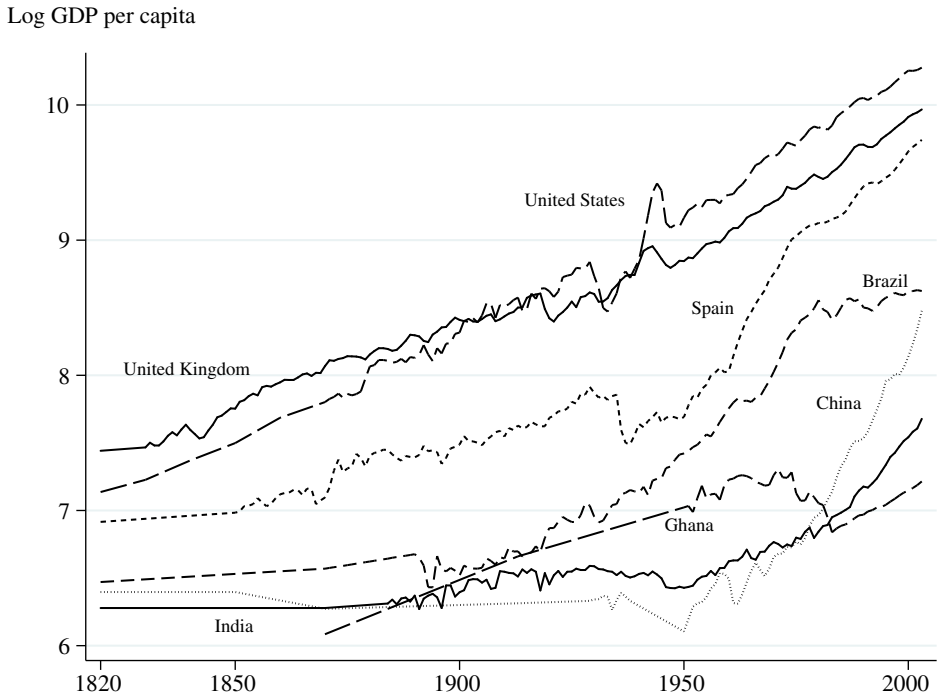

FIGURE 1.12 The evolution of income per capita in the United States, the United Kindgom, Spain, Brazil, China, India, and Ghana, 1820–2000.

1.5 Conditional Convergence

I have so far documented the large differences in income per capita across nations, the slight divergence in economic fortunes over the postwar era, and the much larger divergence since the early 1800s. The analysis focused on the unconditional distribution of income per capita (or per worker). In particular, we looked at whether the income gap between two countries increases or decreases regardless of these countries' characteristics (e.g., institutions, policies, technology, or even investments). Barro and Sala-i-Martin (1991, 1992, 2004) argue that it is instead more informative to look at the conditional distribution. Here the question is whether the income gap between two countries that are similar in observable characteristics is becoming narrower or wider over time. In this case, the picture is one of conditional convergence: in the postwar period, the income gap between countries that share the same characteristics typically closes over time (though it does so quite slowly). This is important both for understanding the statistical properties of the world income distribution and also as an input into the types of theories that we would like to develop.

How do we capture conditional convergence? Consider a typical *Barro growth regression:*

$$g_{i,t,t-1} = \alpha \log y_{i,t-1} + \mathbf{X}_{i,t-1}^{T}\boldsymbol{\beta} + \varepsilon_{i,t}, \tag{1.1}$$

where $g_{i,t,t-1}$ is the annual growth rate between dates $t-1$ and t in country i, $y_{i,t-1}$ is output per worker (or income per capita) at date $t-1$, $\mathbf{X}$ is a vector of other variables included in the regression with coefficient vector $\boldsymbol{\beta}$ ($\mathbf{X}^T$ denotes the transpose of this vector), and $\varepsilon_{i,t}$

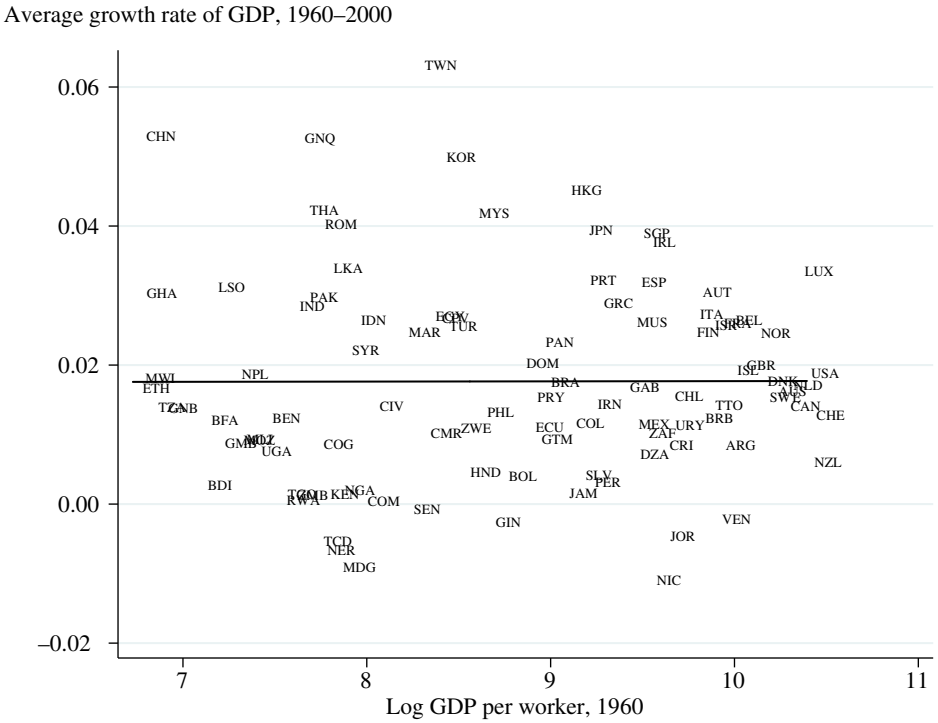

FIGURE 1.13 Annual growth rate of GDP per worker between 1960 and 2000 versus log GDP per worker in 1960 for the entire world.

is an error term capturing all other omitted factors. The variables in $\mathbf{X}$ are included because they are potential determinants of steady-state income and/or growth. First note that without covariates, (1.1) is quite similar to the relationship shown in Figure 1.9. In particular, since $g_{i,t,t-1} \approx \log y_{i,t} - \log y_{i,t-1}$, (1.1) can be written as

$$\log y_{i,t} \approx (1 + \alpha) \log y_{i,t-1} + \varepsilon_{i,t}.$$

Figure 1.9 showed that the relationship between log GDP per worker in 2000 and log GDP per worker in 1960 can be approximated by the 45° line, so that in terms of this equation, α should be approximately equal to 0. This observation is confirmed by Figure 1.13, which depicts the relationship between the (geometric) average growth rate between 1960 and 2000 and log GDP per worker in 1960. This figure reiterates that there is no "unconditional" convergence for the entire world—no tendency for poorer nations to become relatively more prosperous—over the postwar period.

While there is no convergence for the entire world, when we look among the member nations of the Organisation for Economic Co-operation and Development (OECD),[2] we see a different pattern. Figure 1.14 shows that there is a strong negative relationship between log GDP per worker in 1960 and the annual growth rate between 1960 and 2000. What distinguishes this sample from the entire world sample is the relative homogeneity of the OECD countries, which

2. "OECD" here refers to the members that joined the OECD in the 1960s (this excludes Australia, New Zealand, Mexico, and Korea). The figure also excludes Germany because of lack of comparable data after reunification.

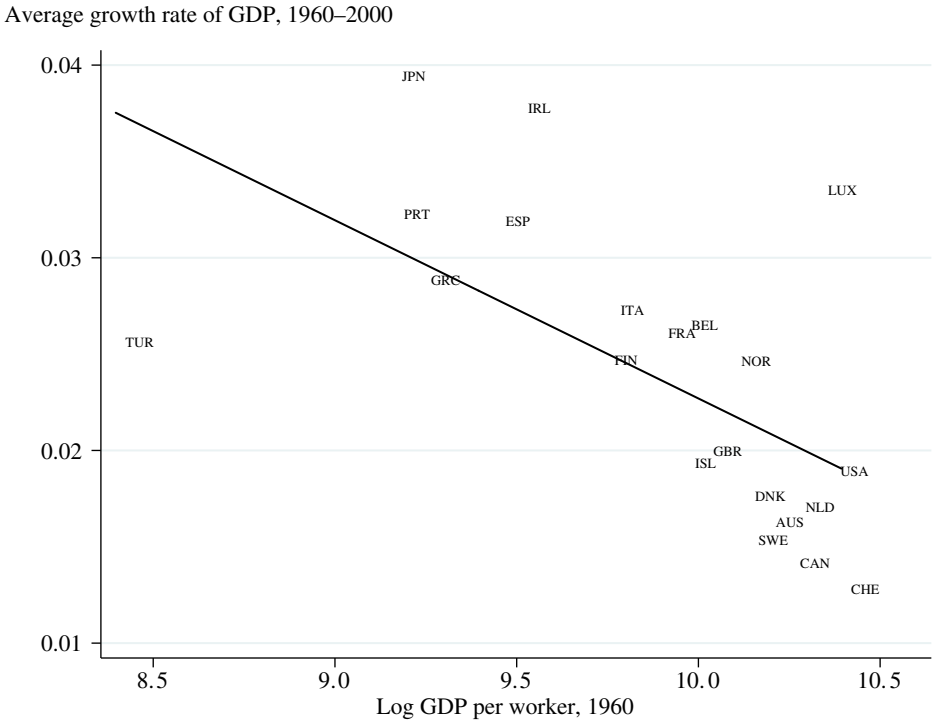

FIGURE 1.14 Annual growth rate of GDP per worker between 1960 and 2000 versus log GDP per worker in 1960 for core OECD countries.

have much more similar institutions, policies, and initial conditions than for the entire world. Thus there might be a type of conditional convergence when we control for certain country characteristics potentially affecting economic growth.

This is what the vector $\mathbf{X}$ captures in (1.1). In particular, when this vector includes such variables as years of schooling or life expectancy, using cross-sectional regressions Barro and Sala-i-Martin estimate α to be approximately -0.02, indicating that the income gap between countries that have the same human capital endowment has been narrowing over the postwar period on average at about 2 percent per year. When this equation is estimated using panel data and the vector $\mathbf{X}$ includes a full set of country fixed effects, the estimates of α become more negative, indicating faster convergence.

In summary, there is no evidence of (unconditional) convergence in the world income distribution over the postwar era (in fact, the evidence suggests some amount of divergence in incomes across nations). But there is some evidence for conditional convergence, meaning that the income gap between countries that are similar in observable characteristics appears to narrow over time. This last observation is relevant both for recognizing among which countries the economic divergence has occurred and for determining what types of models we should consider for understanding the process of economic growth and the differences in economic performance across nations. For example, we will see that many growth models, including the basic Solow and the neoclassical growth models, suggest that there should be transitional dynamics as economies below their steady-state (target) level of income per capita grow toward that level. Conditional convergence is consistent with this type of transitional dynamics.

1.6 Correlates of Economic Growth

The previous section emphasized the importance of certain country characteristics that might be related to the process of economic growth. What types of countries grow more rapidly? Ideally, this question should be answered at a causal level. In other words, we would like to know which specific characteristics of countries (including their policies and institutions) have a causal effect on growth. "Causal effect" refers to the answer to the following counterfactual thought experiment: if, all else being equal, a particular characteristic of the country were changed exogenously (i.e., not as part of equilibrium dynamics or in response to a change in other observable or unobservable variables), what would be the effect on equilibrium growth? Answering such causal questions is quite challenging, precisely because it is difficult to isolate changes in endogenous variables that are not driven by equilibrium dynamics or by omitted factors.

For this reason, let us start with the more modest question of what factors correlate with postwar economic growth. With an eye to the theories to come in the next two chapters, the two obvious candidates to look at are investments in physical and human capital (education).

Figure 1.15 shows a positive association between the average investment to GDP ratio and economic growth between 1960 and 2000. Figure 1.16 shows a positive correlation between average years of schooling and economic growth. These figures therefore suggest that the countries that have grown faster are typically those that have invested more in physical and human capital. It has to be stressed that these figures do not imply that physical or human capital investment are the causes of economic growth (even though we expect from basic economic theory that they should contribute to growth). So far these are simply correlations, and they

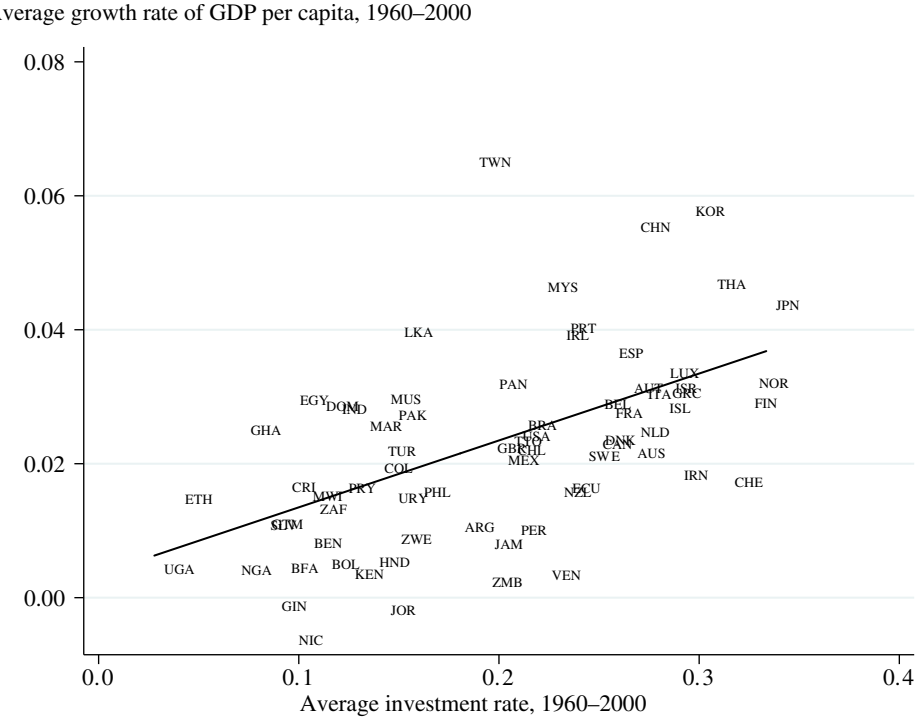

FIGURE 1.15 The relationship between average growth of GDP per capita and average growth of investments to GDP ratio, 1960–2000.

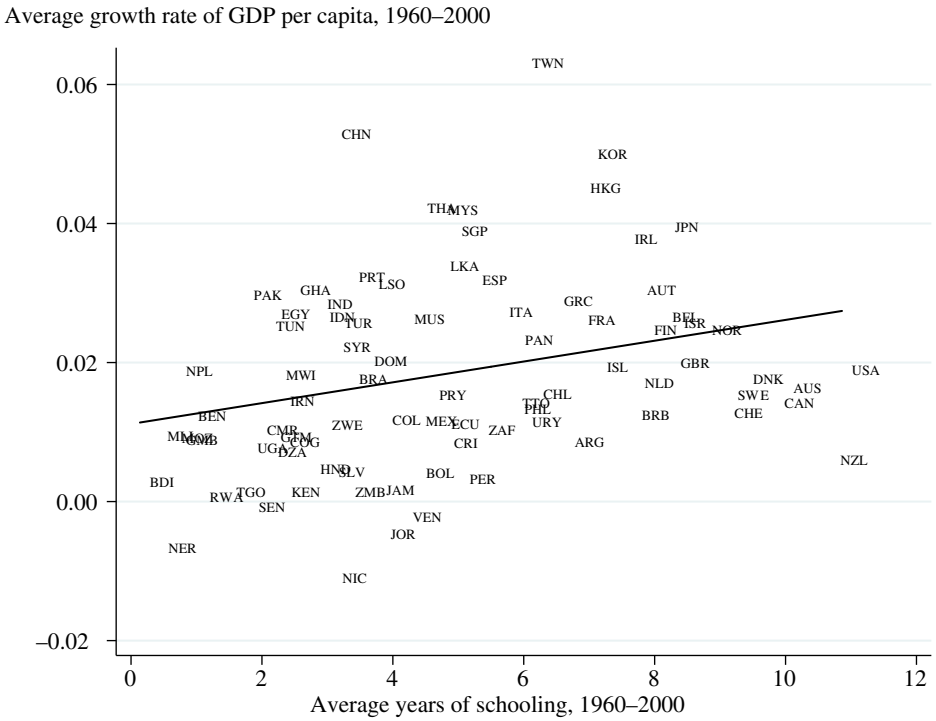

FIGURE 1.16 The relationship between average growth of GDP per capita and average years of schooling, 1960–2000.

are likely driven, at least in part, by omitted factors affecting both investment and schooling on the one hand and economic growth on the other.

We investigate the role of physical and human capital in economic growth further in Chapter 3. One of the major points that emerges from the analysis in Chapter 3 is that focusing only on physical and human capital is not sufficient. Both to understand the process of sustained economic growth and to account for large cross-country differences in income, we also need to understand why societies differ in the efficiency with which they use their physical and human capital. Economists normally use the shorthand expression "technology" to capture factors other than physical and human capital that affect economic growth and performance. It is therefore important to remember that variations in technology across countries include not only differences in production techniques and in the quality of machines used in production but also disparities in productive efficiency (see in particular Chapter 21 on differences in productive efficiency resulting from the organization of markets and from market failures). A detailed study of technology (broadly construed) is necessary for understanding both the worldwide process of economic growth and cross-country differences. The role of technology in economic growth is investigated in Chapter 3 and later chapters.

1.7 From Correlates to Fundamental Causes

The correlates of economic growth, such as physical capital, human capital, and technology, is our first topic of study. But these are only proximate causes of economic growth and economic success (even if we convince ourselves that there is an element of causality in the correlations

shown above). It would not be entirely satisfactory to explain the process of economic growth and cross-country differences with technology, physical capital, and human capital, since presumably there are reasons technology, physical capital, and human capital differ across countries. If these factors are so important in generating cross-country income differences and causing the takeoff into modern economic growth, why do certain societies fail to improve their technologies, invest more in physical capital, and accumulate more human capital?

Let us return to Figure 1.8 to illustrate this point further. This figure shows that South Korea and Singapore have grown rapidly over the past 50 years, while Nigeria has failed to do so. We can try to explain the successful performances of South Korea and Singapore by looking at the proximate causes of economic growth. We can conclude, as many have done, that rapid capital accumulation has been a major cause of these growth miracles and debate the relative roles of human capital and technology. We can simply blame the failure of Nigeria to grow on its inability to accumulate capital and to improve its technology. These perspectives are undoubtedly informative for understanding the mechanics of economic successes and failures of the postwar era. But at some level they do not provide answers to the central questions: How did South Korea and Singapore manage to grow, while Nigeria failed to take advantage of its growth opportunities? If physical capital accumulation is so important, why did Nigeria fail to invest more in physical capital? If education is so important, why are education levels in Nigeria still so low, and why is existing human capital not being used more effectively? The answer to these questions is related to the *fundamental causes* of economic growth—the factors potentially affecting why societies make different technology and accumulation choices.

At some level, fundamental causes are the factors that enable us to link the questions of economic growth to the concerns of the rest of the social sciences and ask questions about the roles of policies, institutions, culture, and exogenous environmental factors. At the risk of oversimplifying complex phenomena, we can think of the following list of potential fundamental causes: (1) luck (or multiple equilibria) that lead to divergent paths among societies with identical opportunities, preferences, and market structures; (2) geographic differences that affect the environment in which individuals live and influence the productivity of agriculture, the availability of natural resources, certain constraints on individual behavior, or even individual attitudes; (3) institutional differences that affect the laws and regulations under which individuals and firms function and shape the incentives they have for accumulation, investment, and trade; and (4) cultural differences that determine individuals' values, preferences, and beliefs. Chapter 4 presents a detailed discussion of the distinction between proximate and fundamental causes and what types of fundamental causes are more promising in explaining the process of economic growth and cross-country income differences.

For now, it is useful to briefly return to the contrast between South Korea and Singapore versus Nigeria and ask the questions (even if we are not in a position to fully answer them yet): Can we say that South Korea and Singapore owe their rapid growth to luck, while Nigeria was unlucky? Can we relate the rapid growth of South Korea and Singapore to geographic factors? Can we relate them to institutions and policies? Can we find a major role for culture? Most detailed accounts of postwar economics and politics in these countries emphasize the role of growth-promoting policies in South Korea and Singapore—including the relative security of property rights and investment incentives provided to firms. In contrast, Nigeria's postwar history is one of civil war, military coups, endemic corruption, and overall an environment that failed to provide incentives to businesses to invest and upgrade their technologies. It therefore seems necessary to look for fundamental causes of economic growth that make contact with these facts. Jumping ahead a little, it already appears implausible that luck can be the major explanation for the differences in postwar economic performance; there were already significant economic differences between South Korea, Singapore, and Nigeria at the beginning of the postwar era. It is also equally implausible to link the divergent fortunes of these countries

to geographic factors. After all, their geographies did not change, but the growth spurts of South Korea and Singapore started in the postwar era. Moreover, even if Singapore benefited from being an island, without hindsight one might have concluded that Nigeria had the best environment for growth because of its rich oil reserves.[3] Cultural differences across countries are likely to be important in many respects, and the rapid growth of many Asian countries is often linked to certain "Asian values." Nevertheless, cultural explanations are also unlikely to adequately explain fundamental causes, since South Korean or Singaporean culture did not change much after the end of World War II, while their rapid growth is a distinctly postwar phenomenon. Moreover, while South Korea grew rapidly, North Korea, whose inhabitants share the same culture and Asian values, has endured one of the most disastrous economic performances of the past 50 years.

This admittedly quick (and partial) account suggests that to develop a better understanding of the fundamental causes of economic growth, we need to look at institutions and policies that affect the incentives to accumulate physical and human capital and improve technology. Institutions and policies were favorable to economic growth in South Korea and Singapore, but not in Nigeria. Understanding the fundamental causes of economic growth is largely about understanding the impact of these institutions and policies on economic incentives and why, for example, they have enhanced growth in South Korea and Singapore but not in Nigeria. The intimate link between fundamental causes and institutions highlighted by this discussion motivates Part VIII, which is devoted to the political economy of growth, that is, to the study of how institutions affect growth and why they differ across countries.

An important caveat should be noted at this point. Discussions of geography, institutions, and culture can sometimes be carried out without explicit reference to growth models or even to growth empirics. After all, this is what many social scientists do outside the field of economics. However, fundamental causes can only have a big impact on economic growth if they affect parameters and policies that have a first-order influence on physical and human capital and technology. Therefore an understanding of the mechanics of economic growth is essential for evaluating whether candidate fundamental causes of economic growth could indeed play the role that is sometimes ascribed to them. Growth empirics plays an equally important role in distinguishing among competing fundamental causes of cross-country income differences. It is only by formulating parsimonious models of economic growth and confronting them with data that we can gain a better understanding of both the proximate and the fundamental causes of economic growth.

1.8 The Agenda

The three major questions that have emerged from the brief discussion are:

1. Why are there such large differences in income per capita and worker productivity across countries?
2. Why do some countries grow rapidly while other countries stagnate?
3. What sustains economic growth over long periods of time, and why did sustained growth start 200 years or so ago?

3. One can turn this reasoning around and argue that Nigeria is poor because of a "natural resource curse," that is, precisely because it has abundant natural resources. But this argument is not entirely compelling, since there are other countries, such as Botswana, with abundant natural resources that have grown rapidly over the past 50 years. More important, the only plausible channel through which abundance of natural resources may lead to worse economic outcomes is related to institutional and political economy factors. Such factors take us to the realm of institutional fundamental causes.

For each question, a satisfactory answer requires a set of well-formulated models that illustrate the mechanics of economic growth and cross-country income differences together with an investigation of the fundamental causes of the different trajectories which these nations have embarked upon. In other words, we need a combination of theoretical models and empirical work.

The traditional growth models—in particular, the basic Solow and the neoclassical models—provide a good starting point, and the emphasis they place on investment and human capital seems consistent with the patterns shown in Figures 1.15 and 1.16. However, we will also see that technological differences across countries (either because of their differential access to technological opportunities or because of differences in the efficiency of production) are equally important. Traditional models treat technology and market structure as given or at best as evolving exogenously (rather like a black box). But if technology is so important, we ought to understand why and how it progresses and why it differs across countries. This motivates our detailed study of endogenous technological progress and technology adoption. Specifically, we will try to understand how differences in technology may arise, persist, and contribute to differences in income per capita. Models of technological change are also useful in thinking about the sources of sustained growth of the world economy over the past 200 years and the reasons behind the growth process that took off 200 years or so ago and has proceeded relatively steadily ever since.

Some of the other patterns encountered in this chapter will inform us about the types of models that have the greatest promise in explaining economic growth and cross-country differences in income. For example, we have seen that cross-country income differences can be accounted for only by understanding why some countries have grown rapidly over the past 200 years while others have not. Therefore we need models that can explain how some countries can go through periods of sustained growth while others stagnate.

Nevertheless, we have also seen that the postwar world income distribution is relatively stable (at most spreading out slightly from 1960 to 2000). This pattern has suggested to many economists that we should focus on models that generate large permanent cross-country differences in income per capita but not necessarily large permanent differences in growth rates (at least not in the recent decades). This argument is based on the following reasoning: with substantially different long-run growth rates (as in models of endogenous growth, where countries that invest at different rates grow at permanently different rates), we should expect significant divergence. We saw above that despite some widening between the top and the bottom, the cross-country distribution of income across the world is relatively stable over the postwar era.

Combining the postwar patterns with the origins of income differences over the past several centuries suggests that we should look for models that can simultaneously account for long periods of significant growth differences and for a distribution of world income that ultimately becomes stationary, though with large differences across countries. The latter is particularly challenging in view of the nature of the global economy today, which allows for the free flow of technologies and large flows of money and commodities across borders. We therefore need to understand how the poor countries fell behind and what prevents them today from adopting and imitating the technologies and the organizations (and importing the capital) of richer nations.

And as the discussion in the previous section suggests, all of these questions can be (and perhaps should be) answered at two distinct but related levels (and in two corresponding steps). The first step is to use theoretical models and data to understand the mechanics of economic growth. This step sheds light on the proximate causes of growth and explains differences in income per capita in terms of differences in physical capital, human capital, and technology,

and these in turn will be related to other variables, such as preferences, technology, market structure, openness to international trade, and economic policies.

The second step is to look at the fundamental causes underlying these proximate factors and investigate why some societies are organized differently than others. Why do societies have different market structures? Why do some societies adopt policies that encourage economic growth while others put up barriers to technological change? These questions are central to a study of economic growth and can only be answered by developing systematic models of the political economy of development and looking at the historical process of economic growth to generate data that can shed light on these fundamental causes.

Our next task is to systematically develop a series of models to understand the mechanics of economic growth. I present a detailed exposition of the mathematical structure of a number of dynamic general equilibrium models that are useful for thinking about economic growth and related macroeconomic phenomena, and I emphasize the implications of these models for the sources of differences in economic performance across societies. Only by understanding these mechanics can we develop a useful framework for thinking about the causes of economic growth and income disparities.

1.9 References and Literature

The empirical material presented in this chapter is largely standard, and parts of it can be found in many books, though interpretations and emphases differ. Excellent introductions, with slightly different emphases, are provided in Jones's (1998, Chapter 1) and Weil's (2005, Chapter 1) undergraduate economic growth textbooks. Barro and Sala-i-Martin (2004) also present a brief discussion of the stylized facts of economic growth, though their focus is on postwar growth and conditional convergence rather than the very large cross-country income differences and the long-run perspective stressed here. Excellent and very readable accounts of the key questions of economic growth, with a similar perspective to the one here, are provided in Helpman (2005) and in Aghion and Howitt's new book (2008). Aghion and Howitt also provide a very useful introduction to many of the same topics discussed in this book.

Much of the data used in this chapter come from Summers-Heston's (Penn World) dataset (latest version, Summers, Heston, and Aten, 2006). These tables are the result of a careful study by Robert Summers and Alan Heston to construct internationally comparable price indices and estimates of income per capita and consumption. PPP adjustment is made possible by these data. Summers and Heston (1991) give a lucid discussion of the methodology for PPP adjustment and its use in the Penn World tables. PPP adjustment enables the construction of measures of income per capita that are comparable across countries. Without PPP adjustment, differences in income per capita across countries can be computed using the current exchange rate or some fundamental exchange rate. There are many problems with such exchange rate–based measures, however. The most important one is that they do not allow for the marked differences in relative prices and even overall price levels across countries. PPP adjustment brings us much closer to differences in real income and real consumption. GDP, consumption, and investment data from the Penn World tables are expressed in 1996 constant U.S. dollars. Information on workers (economically active population), consumption, and investment are also from this dataset. Life expectancy data are from the World Bank's World Development Indicators CD-ROM and refer to the average life expectancy of males and females at birth. This dataset also contains a range of other useful information. Schooling data are from Barro and Lee's (2001) dataset, which contains internationally comparable information on years of schooling. Throughout, cross-country figures use the World Bank labels to denote the identity

of individual countries. A list of the labels can be found in http://unstats.un.org/unsd/methods/m49/m49alpha.htm.

In all figures and regressions, growth rates are computed as geometric averages. In particular, the geometric average growth rate of output per capita y between date t and $t + T$ is

$$g_{t,t+T} \equiv \left(\frac{y_{t+T}}{y_t} \right)^{1/T} - 1.$$

The geometric average growth rate is more appropriate to use in the context of income per capita than is the arithmetic average, since the growth rate refers to proportional growth. It can be easily verified from this formula that if $y_{t+1} = (1 + g)\, y_t$ for all t, then $g_{t+T} = g$.

Historical data are from various works by Angus Maddison, in particular, Maddison (2001, 2003). While these data are not as reliable as the estimates from the Penn World tables, the general patterns they show are typically consistent with evidence from a variety of different sources. Nevertheless, there are points of contention. For example, in Figure 1.11 Maddison's estimates show a slow but relatively steady growth of income per capita in Western Europe starting in 1000. This growth pattern is disputed by some historians and economic historians. A relatively readable account, which strongly disagrees with this conclusion, is provided in Pomeranz (2000), who argues that income per capita in Western Europe and the Yangtze Valley in China were broadly comparable as late as 1800. This view also receives support from recent research by Allen (2004), which documents that the levels of agricultural productivity in 1800 were comparable in Western Europe and China. Acemoglu, Johnson, and Robinson (2002, 2005b) use urbanization rates as a proxy for income per capita and obtain results that are intermediate between those of Maddison and Pomeranz. The data in Acemoglu, Johnson, and Robinson (2002) also confirm that there were very limited income differences across countries as late as the 1500s and that the process of rapid economic growth started in the nineteenth century (or perhaps in the late eighteenth century). Recent research by Broadberry and Gupta (2006) also disputes Pomeranz's arguments and gives more support to a pattern in which there was already an income gap between Western Europe and China by the end of the eighteenth century.

The term "takeoff" used in Section 1.4 is introduced in Walter Rostow's famous book *The Stages of Economic Growth* (1960) and has a broader connotation than the term "industrial revolution," which economic historians typically use to refer to the process that started in Britain at the end of the eighteenth century (e.g., Ashton, 1969). Mokyr (1993) contains an excellent discussion of the debate on whether the beginning of industrial growth was due to a continuous or discontinuous change. Consistent with my emphasis here, Mokyr concludes that this is secondary to the more important fact that the modern process of growth *did* start around this time.

There is a large literature on the correlates of economic growth, starting with Barro (1991). This work is surveyed in Barro and Sala-i-Martin (2004) and Barro (1997). Much of this literature, however, interprets these correlations as causal effects, even when this interpretation is not warranted (see the discussions in Chapters 3 and 4).

Note that Figures 1.15 and 1.16 show the relationship between average investment and average schooling between 1960 and 2000 and economic growth over the same period. The relationship between the growth of investment and economic growth over this time is similar, but there is a much weaker relationship between growth of schooling and economic growth. This lack of association between growth of schooling and growth of output may be for a number of reasons. First, there is considerable measurement error in schooling estimates (see Krueger and Lindahl, 2001). Second, as shown in some of the models discussed later, the main role of human capital may be to facilitate technology adoption, and thus we may expect a stronger

relationship between the level of schooling and economic growth than between the change in schooling and economic growth (see Chapter 10). Finally, the relationship between the level of schooling and economic growth may be partly spurious, in the sense that it may be capturing the influence of some other omitted factors also correlated with the level of schooling; if this is the case, these omitted factors may be removed when we look at changes. While we cannot reach a firm conclusion on these alternative explanations, the strong correlation between average schooling and economic growth documented in Figure 1.16 is interesting in itself.

The narrowing of differences in income per capita in the world economy when countries are weighted by population is explored in Sala-i-Martin (2005). Deaton (2005) contains a critique of Sala-i-Martin's approach. The point that incomes must have been relatively equal around 1800 or before, because there is a lower bound on real incomes necessary for the survival of an individual, was first made by Maddison (1991), and was later popularized by Pritchett (1997). Maddison's estimates of GDP per capita and Acemoglu, Johnson, and Robinson's (2002) estimates based on urbanization confirm this conclusion.

The estimates of the density of income per capita reported in this chapter are similar to those used by Quah (1993, 1997) and Jones (1997). These estimates use a nonparametric Gaussian kernel. The specific details of the kernel estimation do not change the general shape of the densities. Quah was also the first to emphasize the stratification in the world income distribution and the possible shift toward a bimodal distribution, which is visible in Figure 1.3. He dubbed this the "Twin Peaks" phenomenon (see also Durlauf and Quah, 1999). Barro (1991) and Barro and Sala-i-Martin (1992, 2004) emphasize the presence and importance of conditional convergence and argue against the relevance of the stratification pattern emphasized by Quah and others. The estimate of conditional convergence of about 2% per year is from Barro and Sala-i-Martin (1992). Caselli, Esquivel, and Lefort (1996) show that panel data regressions lead to considerably higher rates of conditional convergence.

Marris (1982) and Baumol (1986) were the first economists to conduct cross-country studies of convergence. However, the data at the time were of lower quality than the Summers-Heston data and also were available for only a selected sample of countries. Barro's (1991) and Barro and Sala-i-Martin's (1992) work using the Summers-Heston dataset has been instrumental in generating renewed interest in cross-country growth regressions.

The data on GDP growth and black real wages in South Africa are from Wilson (1972). Wages refer to real wages in gold mines. Feinstein (2005) provides an excellent economic history of South Africa. The implications of the British industrial revolution for real wages and living standards of workers are discussed in Mokyr (1993). Another example of rapid economic growth with falling real wages is provided by the experience of the Mexican economy in the early twentieth century (see Gomez-Galvarriato, 1998). There is also evidence that during this period, the average height of the population might have been declining, which is often associated with falling living standards (see López-Alonso and Porras Condey, 2004).

There is a major debate on the role of technology and capital accumulation in the growth experiences of East Asian nations, particularly South Korea and Singapore. See Young (1991, 1995) for the argument that increases in physical capital and labor inputs explain almost all of the rapid growth in these two countries. See Klenow and Rodriguez (1997) and Hsieh (2002) for the opposite point of view.

The difference between proximate and fundamental causes is discussed further in later chapters. This distinction is emphasized in a different context by Diamond (1997), though it is also implicitly present in North and Thomas's (1973) classic book. It is discussed in detail in the context of long-run economic development and economic growth in Acemoglu, Johnson, and Robinson (2005a). I revisit these issues in greater detail in Chapter 4.

2

The Solow Growth Model

The previous chapter introduced a number of basic facts and posed the main questions concerning the sources of economic growth over time and the causes of differences in economic performance across countries. These questions are central not only to growth theory but also to macroeconomics and the social sciences more generally. Our next task is to develop a simple framework that can help us think about the proximate causes and the mechanics of the process of economic growth and cross-country income differences. We will use this framework both to study potential sources of economic growth and also to perform simple comparative statics to gain an understanding of which country characteristics are conducive to higher levels of income per capita and more rapid economic growth.

Our starting point is the so-called Solow-Swan model named after Robert (Bob) Solow and Trevor Swan, or simply the Solow model, named after the more famous of the two economists. These economists published two pathbreaking articles in the same year, 1956 (Solow, 1956; Swan, 1956) introducing the Solow model. Bob Solow later developed many implications and applications of this model and was awarded the Nobel prize in economics for his contributions. This model has shaped the way we approach not only economic growth but also the entire field of macroeconomics. Consequently, a by-product of our analysis of this chapter is a detailed exposition of a workhorse model of macroeconomics.

The Solow model is remarkable in its simplicity. Looking at it today, one may fail to appreciate how much of an intellectual breakthrough it was. Before the advent of the Solow growth model, the most common approach to economic growth built on the model developed by Roy Harrod and Evsey Domar (Harrod, 1939; Domar, 1946). The Harrod-Domar model emphasized potential dysfunctional aspects of economic growth, for example, how economic growth could go hand-in-hand with increasing unemployment (see Exercise 2.23 on this model). The Solow model demonstrated why the Harrod-Domar model was not an attractive place to start. At the center of the Solow growth model, distinguishing it from the Harrod-Domar model, is the neoclassical aggregate production function. This function not only enables the Solow model to make contact with microeconomics, but as we will see in the next chapter, it also serves as a bridge between the model and the data.

An important feature of the Solow model, which is shared by many models presented in this book, is that it is a simple and abstract representation of a complex economy. At first, it may appear too simple or too abstract. After all, to do justice to the process of growth or macroeconomic equilibrium, we have to consider households and individuals with different tastes, abilities, incomes, and roles in society; various sectors; and multiple social interactions. The Solow model cuts through these complications by constructing a simple one-

good economy, with little reference to individual decisions. Therefore, the Solow model should be thought of as a starting point and a springboard for richer models.

In this chapter, I present the basic Solow model. The closely related neoclassical growth model is presented in Chapter 8.

2.1 The Economic Environment of the Basic Solow Model

Economic growth and development are dynamic processes and thus necessitate dynamic models. Despite its simplicity, the Solow growth model is a dynamic general equilibrium model (though, importantly, many key features of dynamic general equilibrium models emphasized in Chapter 5, such as preferences and dynamic optimization, are missing in this model).

The Solow model can be formulated in either discrete or continuous time. I start with the discrete-time version, because it is conceptually simpler and more commonly used in macroeconomic applications. However, many growth models are formulated in continuous time, and I then provide a detailed exposition of the continuous-time version of the Solow model and show that it is often more convenient to work with.

2.1.1 Households and Production

Consider a closed economy, with a unique final good. The economy is in discrete time running to an infinite horizon, so that time is indexed by $t = 0, 1, 2, \ldots$. Time periods here may correspond to days, weeks, or years. For now, we do not need to specify the time scale.

The economy is inhabited by a large number of households. Throughout the book I use the terms *households, individuals,* and *agents* interchangeably. The Solow model makes relatively few assumptions about households, because their optimization problem is not explicitly modeled. This lack of optimization on the household side is the main difference between the Solow and the *neoclassical growth* models. The latter is the Solow model plus dynamic consumer (household) optimization. To fix ideas, you may want to assume that all households are identical, so that the economy trivially admits *a representative household*—meaning that the demand and labor supply side of the economy can be represented as if it resulted from the behavior of a single household. The representative household assumption is discussed in detail in Chapter 5.

What do we need to know about households in this economy? The answer is: not much. We have not yet endowed households with preferences (utility functions). Instead, for now, households are assumed to save a constant exogenous fraction $s \in (0, 1)$ of their disposable income—regardless of what else is happening in the economy. This assumption is the same as that used in basic Keynesian models and the Harrod-Domar model mentioned above. It is also at odds with reality. Individuals do not save a constant fraction of their incomes; if they did, then an announcement by the government that there will be a large tax increase next year should have no effect on their savings decisions, which seems both unreasonable and empirically incorrect. Nevertheless, the exogenous constant saving rate is a convenient starting point, and we will spend a lot of time in the rest of the book analyzing how consumers behave and make intertemporal choices.

The other key agents in the economy are firms. Firms, like consumers, are highly heterogeneous in practice. Even within a narrowly defined sector of an economy, no two firms are identical. But again for simplicity, let us start with an assumption similar to the representative household assumption, but now applied to firms: suppose that all firms in this economy have access to the same production function for the final good, or that the economy admits a

representative firm, with a representative (or aggregate) production function. The conditions under which this representive firm assumption is reasonable are also discussed in Chapter 5. The aggregate production function for the unique final good is written as

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

where $Y(t)$ is the total amount of production of the final good at time t, $K(t)$ is the capital stock, $L(t)$ is total employment, and $A(t)$ is technology at time t. Employment can be measured in different ways. For example, we may want to think of $L(t)$ as corresponding to hours of employment or to number of employees. The capital stock $K(t)$ corresponds to the quantity of "machines" (or more specifically, equipment and structures) used in production, and it is typically measured in terms of the value of the machines. There are also multiple ways of thinking of capital (and equally many ways of specifying how capital comes into existence). Since the objective here is to start with a simple workable model, I make the rather sharp simplifying assumption that capital is the same as the final good of the economy. However, instead of being consumed, capital is used in the production process of more goods. To take a concrete example, think of the final good as "corn." Corn can be used both for consumption and as an input, as seed, for the production of more corn tomorrow. Capital then corresponds to the amount of corn used as seed for further production.

Technology, on the other hand, has no natural unit, and $A(t)$ is simply a *shifter* of the production function (2.1). For mathematical convenience, I often represent $A(t)$ in terms of a number, but it is useful to bear in mind that, at the end of the day, it is a representation of a more abstract concept. As noted in Chapter 1, we may often want to think of a broad notion of technology, incorporating the effects of the organization of production and of markets on the efficiency with which the factors of production are utilized. In the current model, $A(t)$ represents all these effects.

A major assumption of the Solow growth model (and of the neoclassical growth model we will study in Chapter 8) is that technology is *free:* it is publicly available as a nonexcludable, nonrival good. Recall that a good is *nonrival* if its consumption or use by others does not preclude an individual's consumption or use. It is *nonexcludable*, if it is impossible to prevent another person from using or consuming it. Technology is a good candidate for a nonexcludable, nonrival good; once the society has some knowledge useful for increasing the efficiency of production, this knowledge can be used by any firm without impinging on the use of it by others. Moreover, it is typically difficult to prevent firms from using this knowledge (at least once it is in the public domain and is not protected by patents). For example, once the society knows how to make wheels, everybody can use that knowledge to make wheels without diminishing the ability of others to do the same (thus making the knowledge to produce wheels nonrival). Moreover, unless somebody has a well-enforced patent on wheels, anybody can decide to produce wheels (thus making the knowhow to produce wheels nonexcludable). The implication of the assumptions that technology is nonrival and nonexcludable is that $A(t)$ is freely available to all potential firms in the economy and firms do not have to pay for making use of this technology. Departing from models in which technology is freely available is a major step toward understanding technological progress and will be our focus in Part IV.

As an aside, note that some authors use x_t or K_t when working with discrete time and reserve the notation $x(t)$ or $K(t)$ for continuous time. Since I go back and forth between continuous and discrete time, I use the latter notation throughout. When there is no risk of confusion, I drop the time arguments, but whenever there is the slightest risk of confusion, I err on the side of caution and include the time arguments.

Let us next impose the following standard assumptions on the aggregate production function.

Assumption 1 (Continuity, Differentiability, Positive and Diminishing Marginal Products, and Constant Returns to Scale) *The production function $F : \mathbb{R}^3_+ \to \mathbb{R}_+$ is twice differentiable in K and L, and satisfies*

$$F_K(K, L, A) \equiv \frac{\partial F(K, L, A)}{\partial K} > 0, \quad F_L(K, L, A) \equiv \frac{\partial F(K, L, A)}{\partial L} > 0,$$

$$F_{KK}(K, L, A) \equiv \frac{\partial^2 F(K, L, A)}{\partial K^2} < 0, \quad F_{LL}(K, L, A) \equiv \frac{\partial^2 F(K, L, A)}{\partial L^2} < 0.$$

Moreover, F exhibits constant returns to scale in K and L.

All of the components of Assumption 1 are important. First, the notation $F : \mathbb{R}^3_+ \to \mathbb{R}_+$ implies that the production function takes nonnegative arguments (i.e., $K, L \in \mathbb{R}_+$) and maps to nonnegative levels of output ($Y \in \mathbb{R}_+$). It is natural that the level of capital and the level of employment should be positive. Since A has no natural units, it could have been negative. But there is no loss of generality in restricting it to be positive. The second important aspect of Assumption 1 is that F is a continuous function in its arguments and is also differentiable. There are many interesting production functions that are not differentiable, and some interesting ones that are not even continuous. But working with differentiable functions makes it possible to use differential calculus, and the loss of some generality is a small price to pay for this convenience. Assumption 1 also specifies that marginal products are positive (so that the level of production increases with the amount of inputs); this restriction also rules out some potential production functions and can be relaxed without much complication (see Exercise 2.8). More importantly, Assumption 1 requires that the marginal products of both capital and labor are diminishing, that is, $F_{KK} < 0$ and $F_{LL} < 0$, so that more capital, holding everything else constant, increases output by less and less. And the same applies to labor. This property is sometimes also referred to as "diminishing returns" to capital and labor. The degree of diminishing returns to capital plays a very important role in many results of the basic growth model. In fact, the presence of diminishing returns to capital distinguishes the Solow growth model from its antecedent, the Harrod-Domar model (see Exercise 2.23).

The other important assumption is that of constant returns to scale. Recall that F exhibits *constant returns to scale* in K and L if it is *linearly homogeneous* (homogeneous of degree 1) in these two variables. More specifically:

Definition 2.1 *Let $K \in \mathbb{N}$. The function $g : \mathbb{R}^{K+2} \to \mathbb{R}$ is* homogeneous of degree m *in $x \in \mathbb{R}$ and $y \in \mathbb{R}$ if*

$$g(\lambda x, \lambda y, z) = \lambda^m g(x, y, z) \text{ for all } \lambda \in \mathbb{R}_+ \text{ and } z \in \mathbb{R}^K.$$

It can be easily verified that linear homogeneity implies that the production function F is concave, though not strictly so (see Exercise 2.2). Linearly homogeneous (constant returns to scale) production functions are particularly useful because of the following theorem.

Theorem 2.1 (Euler's Theorem) *Suppose that $g : \mathbb{R}^{K+2} \to \mathbb{R}$ is differentiable in $x \in \mathbb{R}$ and $y \in \mathbb{R}$, with partial derivatives denoted by g_x and g_y, and is homogeneous of degree m in x and y. Then*

$$mg(x, y, z) = g_x(x, y, z)x + g_y(x, y, z)y \text{ for all } x \in \mathbb{R},\ y \in \mathbb{R}, \text{ and } z \in \mathbb{R}^K.$$

Moreover, $g_x(x, y, z)$ and $g_y(x, y, z)$ are themselves homogeneous of degree $m - 1$ in x and y.

Proof. We have that g is differentiable and

$$\lambda^m g(x, y, z) = g(\lambda x, \lambda y, z). \tag{2.2}$$

Differentiate both sides of (2.2) with respect to λ, which gives

$$m\lambda^{m-1} g(x, y, z) = g_x(\lambda x, \lambda y, z)x + g_y(\lambda x, \lambda y, z)y$$

for any λ. Setting $\lambda = 1$ yields the first result. To obtain the second result, differentiate both sides of (2.2) with respect to x:

$$\lambda g_x(\lambda x, \lambda y, z) = \lambda^m g_x(x, y, z).$$

Dividing both sides by λ establishes the desired result. ■

2.1.2 Endowments, Market Structure, and Market Clearing

The previous subsection has specified household behavior and the technology of production. The next step is to specify endowments, that is, the amounts of labor and capital that the economy starts with and who owns these endowments. We will then be in a position to investigate the allocation of resources in this economy. Resources (for a given set of households and production technology) can be allocated in many different ways, depending on the *institutional structure* of the society. Chapters 5–8 discuss how a social planner wishing to maximize a weighted average of the utilities of households might allocate resources, while Part VIII focuses on the allocation of resources favoring individuals who are politically powerful. The more familiar benchmark for the allocation of resources is to assume a specific set of market institutions, in particular, competitive markets. In competitive markets, households and firms act in a price-taking manner and pursue their own objectives, and prices clear markets. Competitive markets are a natural benchmark, and I start by assuming that all goods and factor markets are competitive. This is yet another assumption that is not totally innocuous. For example, both labor and capital markets have imperfections, with certain important implications for economic growth, and monopoly power in product markets plays a major role in Part IV. But these implications can be best appreciated by starting out with the competitive benchmark.

Before investigating trading in competitive markets, let us also specify the ownership of the endowments. Since competitive markets make sense only in the context of an economy with (at least partial) private ownership of assets and the means of production, it is natural to suppose that factors of production are owned by households. In particular, let us suppose that households own all labor, which they supply inelastically. Inelastic supply means that there is some endowment of labor in the economy, for example, equal to the population, $\bar{L}(t)$, and all of it will be supplied regardless of its (rental) price—as long as this price is nonnegative. The labor market clearing condition can then be expressed as:

$$L(t) = \bar{L}(t) \tag{2.3}$$

for all t, where $L(t)$ denotes the demand for labor (and also the level of employment). More generally, this equation should be written in complementary slackness form. In particular, let the rental price of labor or the wage rate at time t be $w(t)$, then the labor market clearing condition takes the form

$$L(t) \leq \bar{L}(t), w(t) \geq 0 \quad \text{and} \quad \left(L(t) - \bar{L}(t)\right) w(t) = 0. \tag{2.4}$$

The complementary slackness formulation ensures that labor market clearing does not happen at a negative wage—or that if labor demand happens to be low enough, employment could be below $\bar{L}(t)$ at zero wage. However, this will not be an issue in most of the models studied in this book, because Assumption 1 and competitive labor markets ensure that wages are strictly positive (see Exercise 2.1). In view of this result, I use the simpler condition (2.3) throughout and denote both labor supply and employment at time t by $L(t)$.

The households also own the capital stock of the economy and rent it to firms. Let us denote the rental price of capital at time t by $R(t)$. The capital market clearing condition is similar to (2.3) and requires the demand for capital by firms to be equal to the supply of capital by households:

$$K(t) = \bar{K}(t),$$

where $\bar{K}(t)$ is the supply of capital by households and $K(t)$ is the demand by firms. Capital market clearing is straightforward to ensure in the class of models analyzed in this book. In particular, it is sufficient that the amount of capital $K(t)$ used in production at time t (from firms' optimization behavior) be consistent with households' endowments and saving behavior.

Let us take households' initial holdings of capital, $K(0) \geq 0$, as given (as part of the description of the environment). For now how this initial capital stock is distributed among the households is not important, since households' optimization decisions are not modeled explicitly and the economy is simply assumed to save a fraction s of its income. When we turn to models with household optimization below, an important part of the description of the environment will be to specify the preferences and the budget constraints of households.

At this point, I could also introduce the price of the final good at time t, say $P(t)$. But there is no need, since there is a choice of a numeraire commodity in this economy, whose price will be normalized to 1. In particular, as discussed in greater detail in Chapter 5, Walras's Law implies that the price of one of the commodities, the numeraire, should be normalized to 1. In fact, throughout I do something stronger and normalize the price of the final good to 1 in all periods. Ordinarily, one cannot choose more than one numeraire—otherwise, one would be fixing the relative price between the numeraires. But as explained in Chapter 5, we can build on an insight by Kenneth Arrow (Arrow, 1964) that it is sufficient to price *securities* (assets) that transfer one unit of consumption from one date (or state of the world) to another. In the context of dynamic economies, this implies that we need to keep track of an *interest rate* across periods, denoted by $r(t)$, which determines intertemporal prices and enables us to normalize the price of the final good to 1 within each period. Naturally we also need to keep track of the wage rate $w(t)$, which determines the price of labor relative to the final good at any date t.

This discussion highlights a central fact: all of the models in this book should be thought of as general equilibrium economies, in which different commodities correspond to the same good at different dates. Recall from basic general equilibrium theory that the same good at different dates (or in different states or localities) is a different commodity. Therefore, in almost all of the models in this book, there will be an infinite number of commodities, since time runs to infinity. This raises a number of special issues, which are discussed in Chapter 5 and later.

Returning to the basic Solow model, the next assumption is that capital depreciates, meaning that machines that are used in production lose some of their value because of wear and tear. In terms of the corn example above, some of the corn that is used as seed is no longer available for consumption or for use as seed in the following period. Let us assume that this depreciation takes an exponential form, which is mathematically very tractable. Thus capital depreciates (exponentially) at the rate $\delta \in (0, 1)$, so that out of 1 unit of capital this period, only $1 - \delta$ is left for next period. Though depreciation here stands for the wear and tear of the machinery, it can also represent the replacement of old machines by new ones in more realistic models (see Chapter 14).

The loss of part of the capital stock affects the interest rate (rate of return on savings) faced by households. Given the assumption of exponential depreciation at the rate δ and the normalization of the price of the final good to 1, the interest rate faced by the households is $r(t) = R(t) - \delta$, where recall that $R(t)$ is the rental price of capital at time t. A unit of final good can be consumed now or used as capital and rented to firms. In the latter case, a household receives $R(t)$ units of good in the next period as the rental price for its savings, but loses δ units of its capital holdings, since δ fraction of capital depreciates over time. Thus the household has given up one unit of commodity dated $t-1$ and receives $1 + r(t) = R(t) + 1 - \delta$ units of commodity dated t, so that $r(t) = R(t) - \delta$. The relationship between $r(t)$ and $R(t)$ explains the similarity between the symbols for the interest rate and the rental rate of capital. The interest rate faced by households plays a central role in the dynamic optimization decisions of households below. In the Solow model, this interest rate does not directly affect the allocation of resources.

2.1.3 Firm Optimization and Equilibrium

We are now in a position to look at the optimization problem of firms and the competitive equilibrium of this economy. Throughout the book I assume that the objective of firms is to maximize profits. Given the assumption that there is an aggregate production function, it is sufficient to consider the problem of a representative firm. Throughout, unless otherwise stated, I also assume that capital markets are functioning, so firms can rent capital in spot markets. For a given technology level $A(t)$, and given factor prices $R(t)$ and $w(t)$, the profit maximization problem of the representative firm at time t can be represented by the following static problem:

$$\max_{K \geq 0, L \geq 0} F(K, L, A(t)) - R(t)K - w(t)L. \tag{2.5}$$

When there are irreversible investments or costs of adjustments, as discussed, for example, in Section 7.8, the maximization problem of firms becomes dynamic. But in the absence of these features, maximizing profits separately at each date t is equivalent to maximizing the net present discounted value of profits. This feature simplifies the analysis considerably.

A couple of additional features are worth noting:

1. The maximization problem is set up in terms of aggregate variables, which, given the representative firm, is without any loss of generality.
2. There is nothing multiplying the F term, since the price of the final good has been normalized to 1. Thus the first term in (2.5) is the revenues of the representative firm (or the revenues of all of the firms in the economy).
3. This way of writing the problem already imposes competitive factor markets, since the firm is taking as given the rental prices of labor and capital, $w(t)$ and $R(t)$ (which are in terms of the numeraire, the final good).
4. This problem is concave, since F is concave (see Exercise 2.2).

An important aspect is that, because F exhibits constant returns to scale (Assumption 1), the maximization problem (2.5) does not have a well-defined solution (see Exercise 2.3); either there does not exist any (K, L) that achieves the maximum value of this program (which is infinity), or $K = L = 0$, or multiple values of (K, L) will achieve the maximum value of this program (when this value happens to be 0). This problem is related to the fact that in a world with constant returns to scale, the size of each individual firm is not determinate (only aggregates are determined). The same problem arises here because (2.5) is written without imposing the condition that factor markets should clear. A competitive equilibrium

requires that all firms (and thus the representative firm) maximize profits and factor markets clear. In particular, the demands for labor and capital must be equal to the supplies of these factors at all times (unless the prices of these factors are equal to zero, which is ruled out by Assumption 1). This observation implies that the representative firm should make zero profits, since otherwise it would wish to hire arbitrarily large amounts of capital and labor exceeding the supplies, which are fixed. It also implies that total demand for labor, L, must be equal to the available supply of labor, $L(t)$. Similarly, the total demand for capital, K, should equal the total supply, $K(t)$. If this were not the case and $L < L(t)$, then there would be an excess supply of labor and the wage would be equal to zero. But this is not consistent with firm maximization, since given Assumption 1, the representative firm would then wish to hire an arbitrarily large amount of labor, exceeding the supply. This argument, combined with the fact that F is differentiable (Assumption 1), implies that given the supplies of capital and labor at time t, $K(t)$ and $L(t)$, factor prices must satisfy the following familiar conditions equating factor prices to marginal products:[1]

$$w(t) = F_L(K(t), L(t), A(t)), \tag{2.6}$$

and

$$R(t) = F_K(K(t), L(t), A(t)). \tag{2.7}$$

Euler's Theorem (Theorem 2.1) then verifies that at the prices (2.6) and (2.7), firms (or the representative firm) make zero profits.

Proposition 2.1 *Suppose Assumption 1 holds. Then, in the equilibrium of the Solow growth model, firms make no profits, and in particular,*

$$Y(t) = w(t)L(t) + R(t)K(t).$$

Proof. This result follows immediately from Theorem 2.1 for the case of constant returns to scale ($m = 1$). ■

Since firms make no profits in equilibrium, the ownership of firms does not need to be specified. All we need to know is that firms are profit-maximizing entities.

In addition to these standard assumptions on the production function, the following boundary conditions, the *Inada conditions,* are often imposed in the analysis of economic growth and macroeconomic equilibria.

Assumption 2 (Inada Conditions) *F satisfies the* Inada conditions

$$\lim_{K\to 0} F_K(K, L, A) = \infty \quad \text{and} \quad \lim_{K\to\infty} F_K(K, L, A) = 0 \text{ for all } L > 0 \text{ and all } A,$$
$$\lim_{L\to 0} F_L(K, L, A) = \infty \quad \text{and} \quad \lim_{L\to\infty} F_L(K, L, A) = 0 \text{ for all } K > 0 \text{ and all } A.$$

Moreover, $F(0, L, A) = 0$ for all L and A.

The role of these conditions—especially in ensuring the existence of *interior equilibria*—will become clear later in this chapter. They imply that the first units of capital and labor

1. An alternative way to derive (2.6) and (2.7) is to consider the cost minimization problem of the representative firm, which takes the form of minimizing $rK + wL$ with respect to K and L, subject to the constraint that $F(K, L, A) = Y$ for some level of output Y. This problem has a unique solution for any given level of Y. Then imposing market clearing, that is, $Y = F(K, L, A)$ with K and L corresponding to the supplies of capital and labor, yields (2.6) and (2.7).

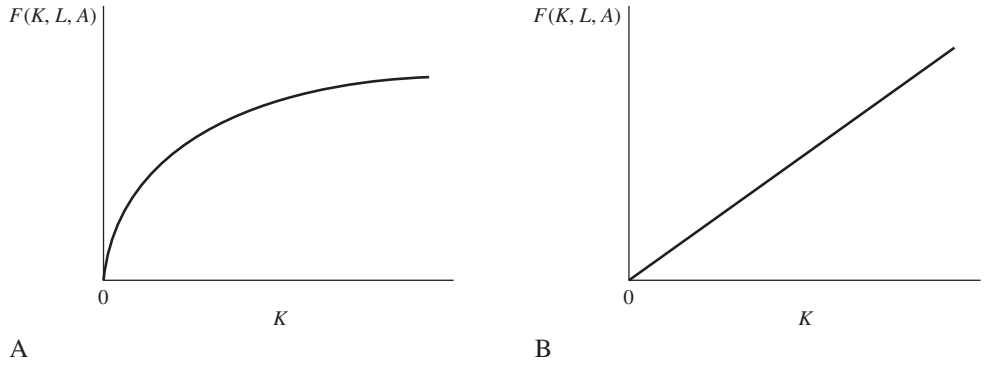

FIGURE 2.1 Production functions. (A) satisfies the Inada conditions in Assumption 2, while (B) does not.

are highly productive and that when capital or labor are sufficiently abundant, their marginal products are close to zero. The condition that $F(0, L, A) = 0$ for all L and A makes capital an essential input. This aspect of the assumption can be relaxed without any major implications for the results in this book. Figure 2.1 shows the production function $F(K, L, A)$ as a function of K, for given L and A, in two different cases; in panel A the Inada conditions are satisfied, while in panel B they are not.

I refer to Assumptions 1 and 2, which can be thought of as the neoclassical technology assumptions, throughout much of the book. For this reason, they are numbered independently from the equations, theorems, and proposition in this chapter.

2.2 The Solow Model in Discrete Time

I next present the dynamics of economic growth in the discrete-time Solow model.

2.2.1 Fundamental Law of Motion of the Solow Model

Recall that K depreciates exponentially at the rate δ, so that the law of motion of the capital stock is given by

$$K(t+1) = (1-\delta)\,K(t) + I(t), \tag{2.8}$$

where $I(t)$ is investment at time t.

From national income accounting for a closed economy, the total amount of final good in the economy must be either consumed or invested, thus

$$Y(t) = C(t) + I(t), \tag{2.9}$$

where $C(t)$ is consumption.[2] Using (2.1), (2.8), and (2.9), any feasible dynamic allocation in this economy must satisfy

$$K(t+1) \leq F(K(t), L(t), A(t)) + (1-\delta)K(t) - C(t)$$

2. In addition, we can introduce government spending $G(t)$ on the right-hand side of (2.9). Government spending does not play a major role in the Solow growth model, thus its introduction is relegated to Exercise 2.7.

for $t = 0, 1, \ldots$. The question is to determine the equilibrium dynamic allocation among the set of feasible dynamic allocations. Here the behavioral rule that households save a constant fraction of their income simplifies the structure of equilibrium considerably (this is a behavioral rule, since it is not derived from the maximization of a well-defined utility function). One implication of this assumption is that any welfare comparisons based on the Solow model have to be taken with a grain of salt, since we do not know what the preferences of the households are.

Since the economy is closed (and there is no government spending), aggregate investment is equal to savings:

$$S(t) = I(t) = Y(t) - C(t).$$

The assumption that households save a constant fraction $s \in (0, 1)$ of their income can be expressed as

$$S(t) = sY(t), \tag{2.10}$$

which, in turn, implies that they consume the remaining $1 - s$ fraction of their income, and thus

$$C(t) = (1 - s)\, Y(t). \tag{2.11}$$

In terms of capital market clearing, (2.10) implies that the supply of capital for time $t + 1$ resulting from households' behavior can be expressed as $K(t + 1) = (1 - \delta)K(t) + S(t) = (1 - \delta)K(t) + sY(t)$. Setting supply and demand equal to each other and using (2.1) and (2.8) yields *the fundamental law of motion* of the Solow growth model:

$$K(t + 1) = sF(K(t), L(t), A(t)) + (1 - \delta)K(t). \tag{2.12}$$

This is a nonlinear difference equation. The equilibrium of the Solow growth model is described by (2.12) together with laws of motion for $L(t)$ and $A(t)$.

2.2.2 Definition of Equilibrium

The Solow model is a mixture of an old-style Keynesian model and a modern dynamic macroeconomic model. Households do not optimize when it comes to their savings or consumption decisions. Instead, their behavior is captured by (2.10) and (2.11). Nevertheless, firms still maximize profits, and factor markets clear. Thus it is useful to start defining equilibria in the way that is customary in modern dynamic macro models.

Definition 2.2 *In the basic Solow model for a given sequence of* $\{L(t), A(t)\}_{t=0}^{\infty}$ *and an initial capital stock* $K(0)$*, an* equilibrium path *is a sequence of capital stocks, output levels, consumption levels, wages, and rental rates* $\{K(t), Y(t), C(t), w(t), R(t)\}_{t=0}^{\infty}$ *such that* $K(t)$ *satisfies (2.12),* $Y(t)$ *is given by (2.1),* $C(t)$ *is given by (2.11), and* $w(t)$ *and* $R(t)$ *are given by (2.6) and (2.7), respectively.*

The most important point to note about Definition 2.2 is that an equilibrium is defined as an entire path of allocations and prices. An economic equilibrium does *not* refer to a static object; it specifies the entire path of behavior of the economy. Note also that Definition 2.2 incorporates the market clearing conditions, (2.6) and (2.7), into the definition of equilibrium. This practice

is standard in macro and growth models. The alternative, which involves describing the equilibrium in more abstract terms, is discussed in Chapter 8 in the context of the neoclassical growth model (see, in particular, Definition 8.1).

2.2.3 Equilibrium without Population Growth and Technological Progress

It is useful to start with the following assumptions, which are relaxed later in this chapter:

1. There is no population growth; total population is constant at some level $L > 0$. Moreover, since households supply labor inelastically, this implies $L(t) = L$.
2. There is no technological progress, so that $A(t) = A$.

Let us define the capital-labor ratio of the economy as

$$k(t) \equiv \frac{K(t)}{L}, \tag{2.13}$$

which is a key object for the analysis. Now using the assumption of constant returns to scale, output (income) per capita, $y(t) \equiv Y(t)/L$, can be expressed as

$$\begin{aligned} y(t) &= F\left(\frac{K(t)}{L}, 1, A\right) \\ &\equiv f(k(t)). \end{aligned} \tag{2.14}$$

In other words, with constant returns to scale, output per capita is simply a function of the capital-labor ratio. Note that $f(k)$ here depends on A, so I could have written $f(k, A)$. I do not do this to simplify the notation and also because until Section 2.7, there will be no technological progress. Thus for now A is constant and can be normalized to $A = 1$.[3] The marginal product and the rental price of capital are then given by the derivative of F with respect to its first argument, which is $f'(k)$. The marginal product of labor and the wage rate are then obtained from Theorem 2.1, so that

$$\begin{aligned} &R(t) = f'(k(t)) > 0 \quad \text{and} \\ &w(t) = f(k(t)) - k(t) f'(k(t)) > 0. \end{aligned} \tag{2.15}$$

The fact that both factor prices are positive follows from Assumption 1, which ensures that the first derivatives of F with respect to capital and labor are always positive.

Example 2.1 (The Cobb-Douglas Production Function) *Let us consider the most common example of production function used in macroeconomics, the Cobb-Douglas production function. I hasten to add the caveat that even though the Cobb-Douglas form is convenient and widely used, it is also very special, and many interesting phenomena discussed later in this book are ruled out by this production function. The Cobb-Douglas production function can be written as*

$$\begin{aligned} Y(t) &= F(K(t), L(t), A(t)) \\ &= AK(t)^{\alpha} L(t)^{1-\alpha}, 0 < \alpha < 1. \end{aligned} \tag{2.16}$$

3. Later, when technological change is taken to be labor-augmenting, the term A can also be taken out, and the per capita production function can be written as $y = Af(k)$, with a slightly different definition of k as effective capital-labor ratio (see, e.g., (2.50) in Section 2.7).

It can easily be verified that this production function satisfies Assumptions 1 and 2, including the constant returns to scale feature imposed in Assumption 1. Dividing both sides by $L(t)$, the per capita production function in (2.14) becomes:

$$y(t) = Ak(t)^{\alpha},$$

where $y(t)$ again denotes output per worker and $k(t)$ is capital-labor ratio as defined in (2.13). The representation of factor prices as in (2.15) can also be verified. From the per capita production function representation, in particular (2.15), the rental price of capital can be expressed as

$$\begin{aligned} R(t) &= \frac{\partial Ak(t)^{\alpha}}{\partial k(t)}, \\ &= \alpha Ak(t)^{-(1-\alpha)}. \end{aligned}$$

Alternatively, in terms of the original production function (2.16), the rental price of capital in (2.7) is given by

$$\begin{aligned} R(t) &= \alpha AK(t)^{\alpha-1}L(t)^{1-\alpha} \\ &= \alpha Ak(t)^{-(1-\alpha)}, \end{aligned}$$

which is equal to the previous expression and thus verifies the form of the marginal product given in (2.15). Similarly, from (2.15),

$$\begin{aligned} w(t) &= Ak(t)^{\alpha} - \alpha Ak(t)^{-(1-\alpha)} \times k(t) \\ &= (1-\alpha)AK(t)^{\alpha}L(t)^{-\alpha}, \end{aligned}$$

which verifies the alternative expression for the wage rate in (2.6).

Returning to the analysis with the general production function, the per capita representation of the aggregate production function enables us to divide both sides of (2.12) by L to obtain the following simple difference equation for the evolution of the capital-labor ratio:

$$k(t+1) = sf(k(t)) + (1-\delta)k(t). \tag{2.17}$$

Since this difference equation is derived from (2.12), it also can be referred to as the *equilibrium difference equation* of the Solow model: it describes the equilibrium behavior of the key object of the model, the capital-labor ratio. The other equilibrium quantities can all be obtained from the capital-labor ratio $k(t)$.

At this point, let us also define a *steady-state equilibrium* for this model.

Definition 2.3 *A* steady-state equilibrium *without technological progress and population growth is an equilibrium path in which $k(t) = k^*$ for all t.*

In a steady-state equilibrium the capital-labor ratio remains constant. Since there is no population growth, this implies that the level of the capital stock will also remain constant. Mathematically, a steady-state equilibrium corresponds to a stationary point of the equilibrium difference equation (2.17). Most of the models in this book admit a steady-state equilibrium. This is also the case for this simple model.

The existence of a steady state can be seen by plotting the difference equation that governs the equilibrium behavior of this economy, (2.17), which is done in Figure 2.2. The thick curve

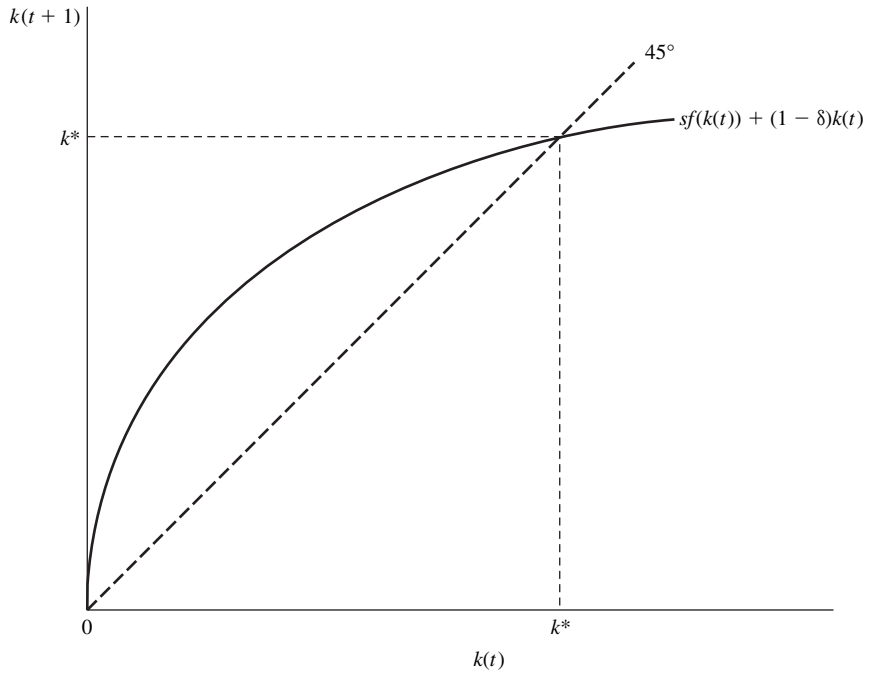

FIGURE 2.2 Determination of the steady-state capital-labor ratio in the Solow model without population growth and technological change.

represents the right-hand side of (2.17) and the dashed line corresponds to the 45° line. Their (positive) intersection gives the steady-state value of the capital-labor ratio k^*, which satisfies

$$\frac{f(k^*)}{k^*} = \frac{\delta}{s}. \tag{2.18}$$

Notice that in Figure 2.2 there is another intersection between (2.17) and the 45° line at $k = 0$. This second intersection occurs because, from Assumption 2, capital is an essential input, and thus $f(0) = 0$. Starting with $k(0) = 0$, there will then be no savings, and the economy will remain at $k = 0$. Nevertheless, I ignore this intersection throughout for a number of reasons. First, $k = 0$ is a steady-state equilibrium only when capital is an essential input and $f(0) = 0$. But as noted above, this assumption can be relaxed without any implications for the rest of the analysis, and when $f(0) > 0$, $k = 0$ is no longer a steady-state equilibrium. This is illustrated in Figure 2.3, which draws (2.17) for the case where $f(0) = \varepsilon$ for some $\varepsilon > 0$. Second, as we will see below, this intersection, even when it exists, is an unstable point; thus the economy would never travel toward this point starting with $K(0) > 0$ (or with $k(0) > 0$). Finally, and most importantly, this intersection holds no economic interest for us.[4]

An alternative visual representation shows the steady state as the intersection between a ray through the origin with slope δ (representing the function δk) and the function $sf(k)$. Figure 2.4, which illustrates this representation, is also useful for two other purposes. First, it depicts the levels of consumption and investment in a single figure. The vertical distance between the horizontal axis and the δk line at the steady-state equilibrium gives the amount of

4. Hakenes and Irmen (2006) show that even with $f(0) = 0$, the Inada conditions imply that in the continuous-time version of the Solow model $k = 0$ may not be the only equilibrium and the economy may move away from $k = 0$.

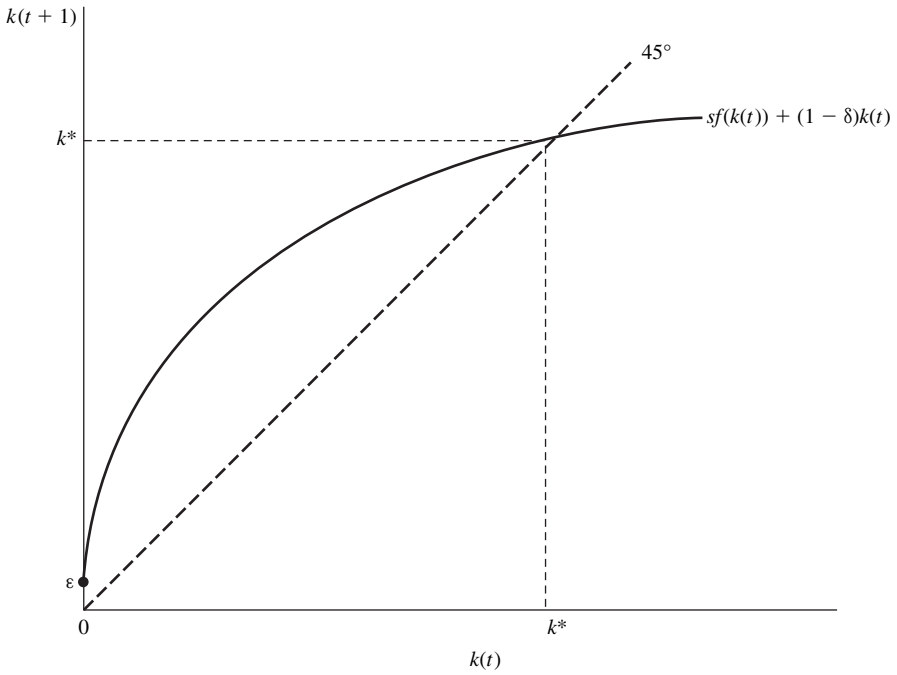

FIGURE 2.3 Unique steady state in the basic Solow model when $f(0) = \varepsilon > 0$.

investment per capita at the steady-state equilibrium (equal to δk^*), while the vertical distance between the function $f(k)$ and the δk line at k^* gives the level of consumption per capita. Clearly, the sum of these two terms make up $f(k^*)$. Second, Figure 2.4 also emphasizes that the steady-state equilibrium in the Solow model essentially sets investment, $sf(k)$, equal to the amount of capital that needs to be replenished, δk. This interpretation is particularly useful when population growth and technological change are incorporated.

This analysis therefore leads to the following proposition (with the convention that the intersection at $k = 0$ is being ignored even though $f(0) = 0$).

Proposition 2.2 *Consider the basic Solow growth model and suppose that Assumptions 1 and 2 hold. Then there exists a unique steady-state equilibrium where the capital-labor ratio* $k^* \in (0, \infty)$ *satisfies (2.18), per capita output is given by*

$$y^* = f(k^*), \tag{2.19}$$

and per capita consumption is given by

$$c^* = (1 - s)\, f(k^*). \tag{2.20}$$

Proof. The preceding argument establishes that any k^* that satisfies (2.18) is a steady state. To establish existence, note that from Assumption 2 (and from l'Hôpital's Rule, see Theorem A.21 in Appendix A), $\lim_{k\to 0} f(k)/k = \infty$ and $\lim_{k\to\infty} f(k)/k = 0$. Moreover, $f(k)/k$ is continuous from Assumption 1, so by the Intermediate Value Theorem (Theorem A.3) there exists k^* such that (2.18) is satisfied. To see uniqueness, differentiate $f(k)/k$ with respect to k, which gives

$$\frac{\partial (f(k)/k)}{\partial k} = \frac{f'(k)k - f(k)}{k^2} = -\frac{w}{k^2} < 0, \tag{2.21}$$

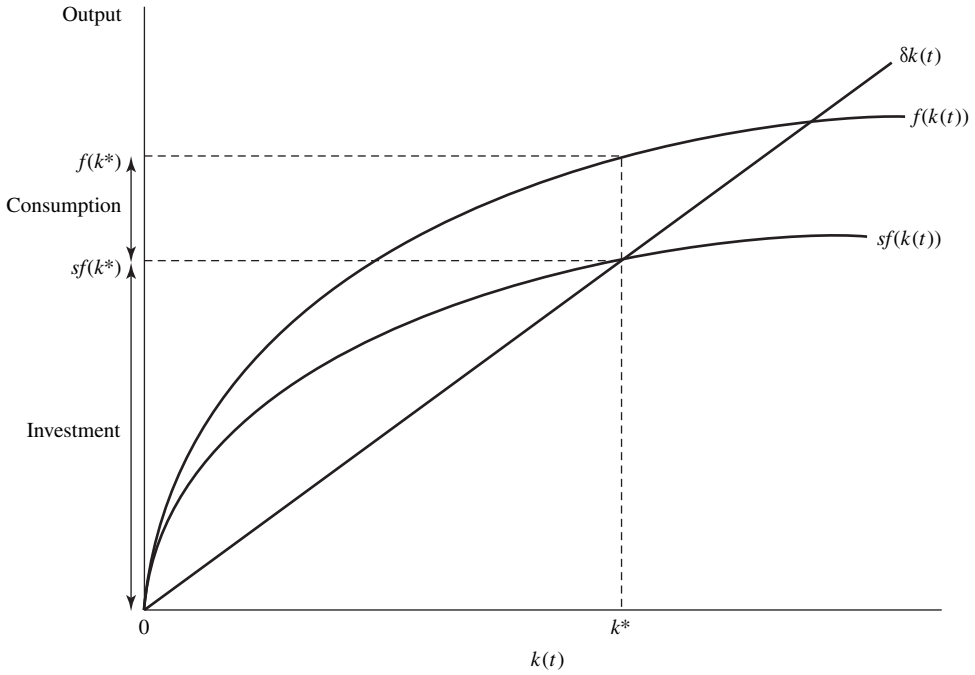

FIGURE 2.4 Investment and consumption in the steady-state equilibrium.

where the last equality in (2.21) uses (2.15). Since $f(k)/k$ is everywhere (strictly) decreasing, there can only exist a unique value k^* that satisfies (2.18). Equations (2.19) and (2.20) then follow by definition. ■

Through a series of examples, Figure 2.5 shows why Assumptions 1 and 2 cannot be dispensed with for establishing the existence and uniqueness results in Proposition 2.2. In the first two panels, the failure of Assumption 2 leads to a situation in which there is no steady-state equilibrium with positive activity, while in the third panel, the failure of Assumption 1 leads to nonuniqueness of steady states.

So far the model is very parsimonious: it does not have many parameters and abstracts from many features of the real world. An understanding of how cross-country differences in certain parameters translate into differences in growth rates or output levels is essential for our focus. This connection will be made in the next proposition. But before doing so, let us generalize the production function in one simple way and assume that

$$f(k) = A\tilde{f}(k),$$

where $A > 0$, so that A is a shift parameter, with greater values corresponding to greater productivity of factors. This type of productivity is referred to as "Hicks-neutral" (see below). For now, it is simply a convenient way of parameterizing productivity differences across countries. Since $f(k)$ satisfies the regularity conditions imposed above, so does $\tilde{f}(k)$.

Proposition 2.3 *Suppose Assumptions 1 and 2 hold and $f(k) = A\tilde{f}(k)$. Denote the steady-state level of the capital-labor ratio by $k^*(A, s, \delta)$ and the steady-state level of output by*

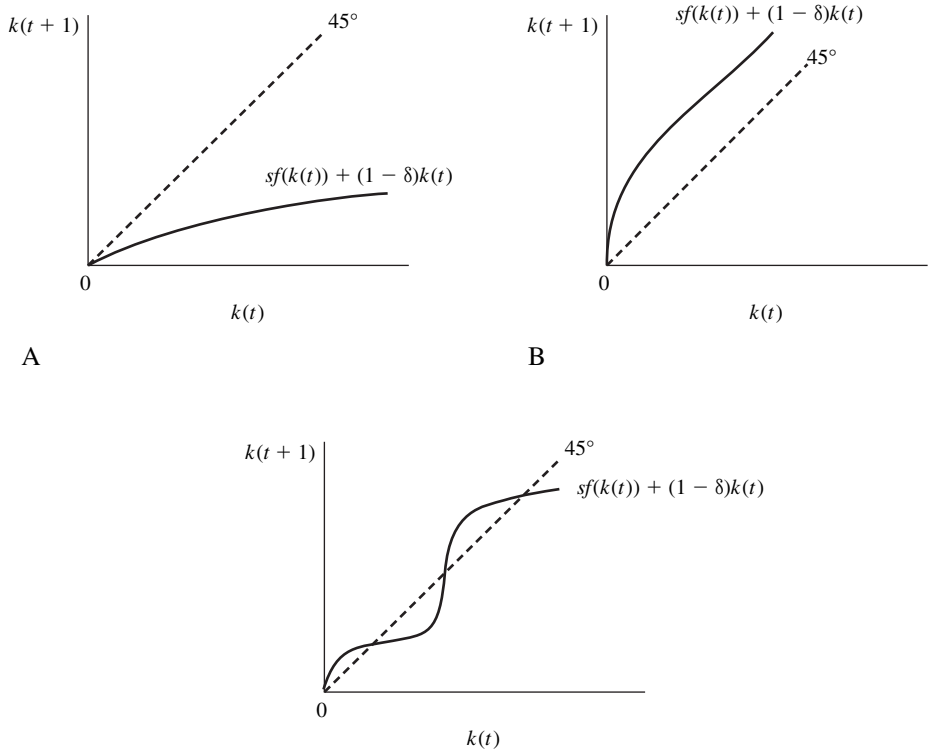

FIGURE 2.5 Examples of nonexistence and nonuniqueness of interior steady states when Assumptions 1 and 2 are not satisfied.

$y^*(A, s, \delta)$ *when the underlying parameters are* A, s, *and* δ. *Then*

$$\frac{\partial k^*(A, s, \delta)}{\partial A} > 0, \quad \frac{\partial k^*(A, s, \delta)}{\partial s} > 0, \quad \text{and } \frac{\partial k^*(A, s, \delta)}{\partial \delta} < 0;$$
$$\frac{\partial y^*(A, s, \delta)}{\partial A} > 0, \quad \frac{\partial y^*(A, s, \delta)}{\partial s} > 0, \quad \text{and } \frac{\partial y^*(A, s, \delta)}{\partial \delta} < 0.$$

Proof. The proof follows immediately by writing

$$\frac{\tilde{f}(k^*)}{k^*} = \frac{\delta}{As},$$

which holds for an open set of values of k^*, A, s, and δ. Now apply the Implicit Function Theorem (Theorem A.25) to obtain the results. For example,

$$\frac{\partial k^*}{\partial s} = \frac{\delta (k^*)^2}{s^2 w^*} > 0,$$

where $w^* = f(k^*) - k^* f'(k^*) > 0$. The other results follow similarly. ■

Therefore countries with higher saving rates and better technologies will have higher capital-labor ratios and will be richer. Those with greater (technological) depreciation will tend to have lower capital-labor ratios and will be poorer. All of the results in Proposition 2.3 are intuitive, and they provide us with a first glimpse of the potential determinants of the capital-labor ratios and output levels across countries.

The same comparative statics with respect to A and δ also apply to c^*. However, it is straightforward to see that c^* is not monotone in the saving rate (e.g., think of the extreme case where $s = 1$). In fact, there exists a unique saving rate, s_{gold}, referred to as the "golden rule" saving rate, which maximizes the steady-state level of consumption. Since we are treating the saving rate as an exogenous parameter and have not specified the objective function of households yet, we cannot say whether the golden rule saving rate is better than some other saving rate. It is nevertheless interesting to characterize what this golden rule saving rate corresponds to. To do this, let us first write the steady-state relationship between c^* and s and suppress the other parameters:

$$\begin{aligned} c^*(s) &= (1-s) f(k^*(s)) \\ &= f(k^*(s)) - \delta k^*(s), \end{aligned}$$

where the second equality exploits the fact that in steady state, $sf(k) = \delta k$. Now differentiating this second line with respect to s (again using the Implicit Function Theorem), we obtain

$$\frac{\partial c^*(s)}{\partial s} = [f'(k^*(s)) - \delta] \frac{\partial k^*}{\partial s}. \tag{2.22}$$

Let us define the golden rule saving rate s_{gold} to be such that $\partial c^*(s_{\text{gold}})/\partial s = 0$. The corresponding steady-state golden rule capital stock is defined as k^*_{gold}. These quantities and the relationship between consumption and the saving rate are plotted in Figure 2.6. The next proposition shows that s_{gold} and k^*_{gold} are uniquely defined.

Proposition 2.4 *In the basic Solow growth model, the highest level of steady-state consumption is reached for s_{gold}, with the corresponding steady-state capital level k^*_{gold} such that*

$$f'(k^*_{\text{gold}}) = \delta. \tag{2.23}$$

Proof. By definition $\partial c^*(s_{\text{gold}})/\partial s = 0$. From Proposition 2.3, $\partial k^*/\partial s > 0$; thus (2.22) can be equal to zero only when $f'(k^*(s_{\text{gold}})) = \delta$. Moreover, when $f'(k^*(s_{\text{gold}})) = \delta$, it can be verified that $\partial^2 c^*(s_{\text{gold}})/\partial s^2 < 0$, so $f'(k^*(s_{\text{gold}})) = \delta$ indeed corresponds to a local maximum. That $f'(k^*(s_{\text{gold}})) = \delta$ also yields the global maximum is a consequence of the following observations: for all $s \in [0, 1]$, we have $\partial k^*/\partial s > 0$, and moreover, when $s < s_{\text{gold}}$, $f'(k^*(s)) - \delta > 0$ by the concavity of f, so $\partial c^*(s)/\partial s > 0$ for all $s < s_{\text{gold}}$. By the converse argument, $\partial c^*(s)/\partial s < 0$ for all $s > s_{\text{gold}}$. Therefore only s_{gold} satisfies $f'(k^*(s)) = \delta$ and gives the unique global maximum of consumption per capita. ■

In other words, there exists a unique saving rate, s_{gold}, and also a unique corresponding capital-labor ratio, k^*_{gold}, given by (2.23), that maximize the level of steady-state consumption. When the economy is below k^*_{gold}, a higher saving rate will increase consumption, whereas when the economy is above k^*_{gold}, steady-state consumption can be raised by saving less. In the latter case, lower savings translate into higher consumption, because the capital-labor ratio of the economy is too high; households are investing too much and not consuming enough. This is the essence of the phenomenon of *dynamic inefficiency,* discussed in greater detail in Chapter 9. For now, recall that there is no explicit utility function here, so statements about inefficiency

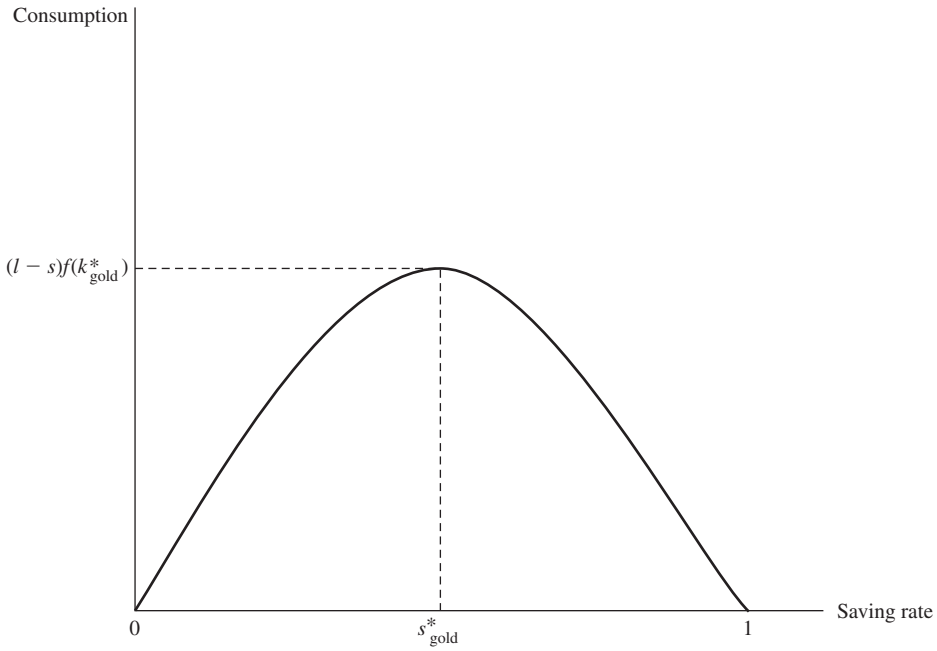

FIGURE 2.6 The golden rule level of saving rate, which maximizes steady-state consumption.

must be considered with caution. In fact, the reason this type of dynamic inefficiency does not generally apply when consumption-saving decisions are endogenized may already be apparent to many of you.

2.3 Transitional Dynamics in the Discrete-Time Solow Model

Proposition 2.2 establishes the existence of a unique steady-state equilibrium (with positive activity). Recall that an equilibrium path does not refer simply to the steady state but to the entire path of capital stock, output, consumption, and factor prices. This is an important point to bear in mind, especially since the term "equilibrium" is used differently in economics than in other disciplines. Typically, in engineering and the physical sciences, an equilibrium refers to a point of rest of a dynamical system, thus to what I have so far referred to as "the steady-state equilibrium." One may then be tempted to say that the system is in "disequilibrium" when it is away from the steady state. However, in economics, the non-steady-state behavior of an economy is also governed by market clearing and optimizing behavior of households and firms. Most economies spend much of their time in non-steady-state situations. Thus we are typically interested in the entire dynamic equilibrium path of the economy, not just in its steady state.

To determine what the equilibrium path of our simple economy looks like, we need to study the transitional dynamics of the equilibrium difference equation (2.17) starting from an arbitrary capital-labor ratio, $k(0) > 0$. Of special interest are the answers to the questions of whether the economy will tend to this steady state starting from an arbitrary capital-labor ratio and how it will behave along the transition path. Recall that the total amount of capital at the beginning of the economy, $K(0) > 0$, is taken as a state variable, while for now, the supply of labor L is fixed. Therefore at time $t = 0$, the economy starts with an arbitrary capital-labor ratio $k(0) = K(0)/L > 0$ as its initial value and then follows the law of motion given by the

difference equation (2.17). Thus the question is whether (2.17) will take us to the unique steady state starting from an arbitrary initial capital-labor ratio.

Before answering this question, recall some definitions and key results from the theory of dynamical systems. Appendix B provides more details and a number of further results. Consider the nonlinear system of autonomous difference equations,

$$\mathbf{x}(t+1) = \mathbf{G}(\mathbf{x}(t)), \tag{2.24}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ and $\mathbf{G} : \mathbb{R}^n \to \mathbb{R}^n$ (where $n \in \mathbb{R}$). Let $\mathbf{x}^*$ be a *fixed point* of the mapping $\mathbf{G}(\cdot)$, that is,

$$\mathbf{x}^* = \mathbf{G}(\mathbf{x}^*).$$

I refer to $\mathbf{x}^*$ as a "steady state" of the difference equation (2.24).[5] The relevant notion of stability is introduced in the next definition.

Definition 2.4 *A steady state* $\mathbf{x}^*$ *is* locally asymptotically stable *if there exists an open set* $B(\mathbf{x}^*)$ *containing* $\mathbf{x}^*$ *such that for any solution* $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ *to (2.24) with* $\mathbf{x}(0) \in B(\mathbf{x}^*)$, $\mathbf{x}(t) \to \mathbf{x}^*$. *Moreover,* $\mathbf{x}^*$ *is* globally asymptotically stable *if for all* $\mathbf{x}(0) \in \mathbb{R}^n$, *for any solution* $\{\mathbf{x}(t)\}_{t=0}^{\infty}$, $\mathbf{x}(t) \to \mathbf{x}^*$.

The next theorem provides the main results on the stability properties of systems of linear difference equations. The following theorems are special cases of the results presented in Appendix B.

Theorem 2.2 (Stability for Systems of Linear Difference Equations) *Consider the following linear difference equation system:*

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}, \tag{2.25}$$

with initial value $\mathbf{x}(0)$, *where* $\mathbf{x}(t) \in \mathbb{R}^n$ *for all* t, $\mathbf{A}$ *is an* $n \times n$ *matrix, and* $\mathbf{b}$ *is a* $n \times 1$ *column vector. Let* $\mathbf{x}^*$ *be the steady state of the difference equation given by* $\mathbf{A}\mathbf{x}^* + \mathbf{b} = \mathbf{x}^*$. *Suppose that all of the eigenvalues of* $\mathbf{A}$ *are strictly inside the unit circle in the complex plane. Then the steady state of the difference equation (2.25),* $\mathbf{x}^*$, *is globally (asymptotically) stable, in the sense that starting from any* $\mathbf{x}(0) \in \mathbb{R}^n$, *the unique solution* $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ *satisfies* $\mathbf{x}(t) \to \mathbf{x}^*$.

Unfortunately, much less can be said about nonlinear systems, but the following is a standard local stability result.

Theorem 2.3 (Local Stability for Systems of Nonlinear Difference Equations) *Consider the following nonlinear autonomous system:*

$$\mathbf{x}(t+1) = \mathbf{G}(\mathbf{x}(t)), \tag{2.26}$$

with initial value $\mathbf{x}(0)$, *where* $\mathbf{G} : \mathbb{R}^n \to \mathbb{R}^n$. *Let* $\mathbf{x}^*$ *be a steady state of this system, that is,* $\mathbf{G}(\mathbf{x}^*) = \mathbf{x}^*$, *and suppose that* $\mathbf{G}$ *is differentiable at* $\mathbf{x}^*$. *Define*

$$\mathbf{A} \equiv D\mathbf{G}(\mathbf{x}^*),$$

5. Various other terms are used to describe $\mathbf{x}^*$, for example, "equilibrium point" or "critical point." Since these other terms have different meanings in economics, I refer to $\mathbf{x}^*$ as a steady state throughout.

where $D\mathbf{G}$ denotes the matrix of partial derivatives (Jacobian) of $\mathbf{G}$. Suppose that all of the eigenvalues of $\mathbf{A}$ are strictly inside the unit circle. Then the steady state of the difference equation (2.26), $\mathbf{x}^*$, is locally (asymptotically) stable, in the sense that there exists an open neighborhood of $\mathbf{x}^*$, $\mathbf{B}(\mathbf{x}^*) \subset \mathbb{R}^n$, such that starting from any $\mathbf{x}(0) \in \mathbf{B}(\mathbf{x}^*)$, $\mathbf{x}(t) \to \mathbf{x}^*$.

An immediate corollary of Theorem 2.3 is the following useful result.

Corollary 2.1

1. *Let $x(t), a, b \in \mathbb{R}$. If $|a| < 1$, then the unique steady state of the linear difference equation $x(t+1) = ax(t) + b$ is globally (asymptotically) stable, in the sense that $x(t) \to x^* = b/(1-a)$.*
2. *Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable in the neighborhood of the steady state x^*, defined by $g(x^*) = x^*$, and suppose that $\left|g'(x^*)\right| < 1$. Then the steady state x^* of the nonlinear difference equation $x(t+1) = g(x(t))$ is locally (asymptotically) stable. Moreover, if g is continuously differentiable and satisfies $\left|g'(x)\right| < 1$ for all $x \in \mathbb{R}$, then x^* is globally (asymptotically) stable.*

Proof. The first part follows immediately from Theorem 2.2. The local stability of g in the second part follows from Theorem 2.3. Global stability follows since

$$\begin{aligned} |x(t+1) - x^*| &= |g(x(t)) - g(x^*)| \\ &= \left| \int_{x^*}^{x(t)} g'(x)dx \right| \\ &< |x(t) - x^*|, \end{aligned}$$

where the second line follows from the Fundamental Theorem of Calculus (Theorem B.2 in Appendix B), and the last inequality uses the hypothesis that $\left|g'(x)\right| < 1$ for all $x \in \mathbb{R}$. This implies that for any $x(0) < x^*$, $\{x(t)\}_{t=0}^{\infty}$ is an increasing sequence. Since $\left|g'(x)\right| < 1$, there cannot exist $x' \neq x^*$ such that $x' = g(x')$, and moreover $\{x(t)\}_{t=0}^{\infty}$ is bounded above by x^*. It therefore converges to x^*. The argument for the case where $x(0) > x^*$ is identical. ■

We can now apply Corollary 2.1 to the equilibrium difference equation (2.17) of the Solow model to establish the local stability of the steady-state equilibrium. Global stability does not directly follow form Corollary 2.1 (since the equivalent of $\left|g'(x)\right| < 1$ for all x is not true), but a slightly different argument can be used to prove this property.

Proposition 2.5 *Suppose that Assumptions 1 and 2 hold. Then the steady-state equilibrium of the Solow growth model described by the difference equation (2.17) is globally asymptotically stable, and starting from any $k(0) > 0$, $k(t)$ monotonically converges to k^*.*

Proof. Let $g(k) \equiv sf(k) + (1-\delta)k$. First observe that $g'(k)$ exists and is always strictly positive, that is, $g'(k) > 0$ for all k. Next, from (2.17),

$$k(t+1) = g(k(t)), \tag{2.27}$$

with a unique steady state at k^*. From (2.18), the steady-state capital k^* satisfies $\delta k^* = sf(k^*)$, or

$$k^* = g(k^*). \tag{2.28}$$

Now recall that $f(\cdot)$ is concave and differentiable from Assumption 1 and satisfies $f(0) = 0$ from Assumption 2. For any strictly concave differentiable function, we have (recall Fact A.23 in Appendix A):

$$f(k) > f(0) + kf'(k) = kf'(k). \tag{2.29}$$

Since (2.29) implies that $\delta = sf(k^*)/k^* > sf'(k^*)$, we have $g'(k^*) = sf'(k^*) + 1 - \delta < 1$. Therefore

$$g'(k^*) \in (0, 1).$$

Corollary 2.1 then establishes local asymptotic stability.

To prove global stability, note that for all $k(t) \in (0, k^*)$,

$$\begin{aligned} k(t+1) - k^* &= g(k(t)) - g(k^*) \\ &= -\int_{k(t)}^{k^*} g'(k)dk, \\ &< 0, \end{aligned}$$

where the first line follows by subtracting (2.28) from (2.27), the second line again uses the Fundamental Theorem of Calculus (Theorem B.2), and the third line follows from the observation that $g'(k) > 0$ for all k. Next, (2.17) also implies

$$\begin{aligned} \frac{k(t+1) - k(t)}{k(t)} &= s\frac{f(k(t))}{k(t)} - \delta \\ &> s\frac{f(k^*)}{k^*} - \delta \\ &= 0, \end{aligned}$$

where the second line uses the fact that $f(k)/k$ is decreasing in k (from (2.29)) and the last line uses the definition of k^*. These two arguments together establish that for all $k(t) \in (0, k^*)$, $k(t+1) \in (k(t), k^*)$. Therefore $\{k(t)\}_{t=0}^{\infty}$ is monotonically increasing and is bounded above by k^*. Moreover, since k^* is the unique steady state (with $k > 0$), there exists no $k' \in (0, k^*)$ such that $k(t+1) = k(t) = k'$ for any t. Therefore $\{k(t)\}_{t=0}^{\infty}$ must monotonically converge to k^*. An identical argument implies that for all $k(t) > k^*$, $k(t+1) \in (k^*, k(t))$ and establishes monotonic convergence starting from $k(0) > k^*$. This completes the proof of global stability. ■

This stability result can be seen diagrammatically in Figure 2.7. Starting from initial capital stock $k(0) > 0$, which is below the steady-state level k^*, the economy grows toward k^* and experiences *capital deepening*—meaning that the capital-labor ratio increases. Together with capital deepening comes growth of per capita income. If instead the economy were to start with $k'(0) > k^*$, it would reach the steady state by decumulating capital and contracting (i.e., by experiencing negative growth).

The following proposition is an immediate corollary of Proposition 2.5.

Proposition 2.6 *Suppose that Assumptions 1 and 2 hold, and $k(0) < k^*$. Then $\{w(t)\}_{t=0}^{\infty}$ is an increasing sequence, and $\{R(t)\}_{t=0}^{\infty}$ is a decreasing sequence. If $k(0) > k^*$, the opposite results apply.*

Proof. See Exercise 2.9. ■

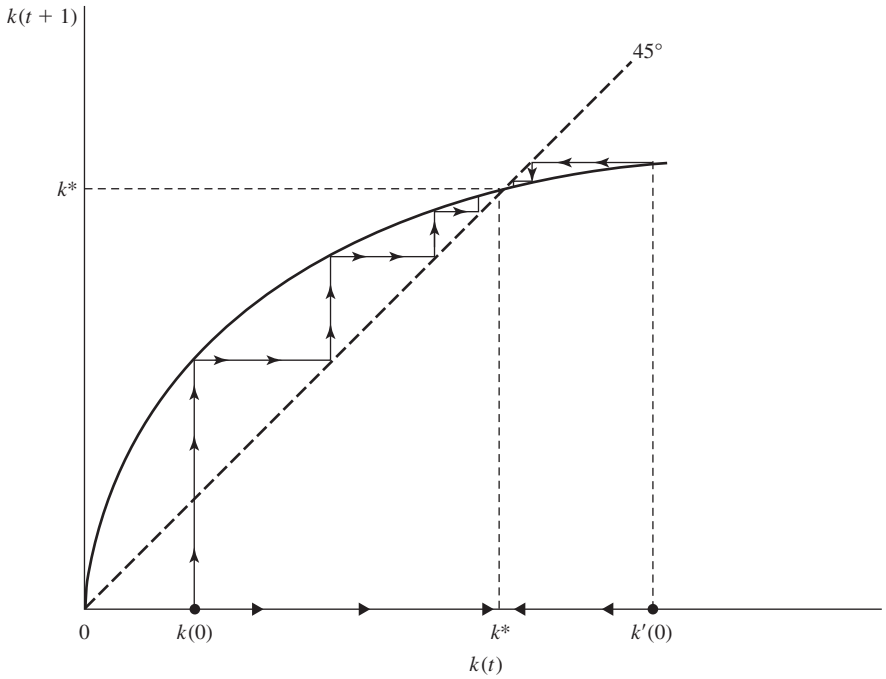

FIGURE 2.7 Transitional dynamics in the basic Solow model.

Recall that when the economy starts with too little capital relative to its labor supply, the capital-labor ratio will increase. Thus the marginal product of capital will fall due to diminishing returns to capital and the wage rate will increase. Conversely, if it starts with too much capital, it will decumulate capital, and in the process the wage rate will decline and the rate of return to capital will increase.

The analysis has established that the Solow growth model has a number of nice properties: unique steady state, global (asymptotic) stability, and finally, simple and intuitive comparative statics. Yet so far it has no growth. The steady state is the point at which there is no growth in the capital-labor ratio, no more capital deepening, and no growth in output per capita. Consequently, the basic Solow model (without technological progress) can only generate economic growth along the transition path to the steady state (starting with $k(0) < k^*$). However this growth is not sustained: it slows down over time and eventually comes to an end. Section 2.7 shows that the Solow model can incorporate economic growth by allowing exogenous technological change. Before doing this, it is useful to look at the relationship between the discrete- and continuous-time formulations.

2.4 The Solow Model in Continuous Time

2.4.1 From Difference to Differential Equations

Recall that the time periods $t = 0, 1, \ldots$ can refer to days, weeks, months, or years. In some sense, the time unit is not important. This arbitrariness suggests that perhaps it may be more convenient to look at dynamics by making the time unit as small as possible, that is, by going to continuous time. While much of modern macroeconomics (outside of growth theory) uses

discrete-time models, many growth models are formulated in continuous time. The continuous-time setup has a number of advantages, since some pathological results of discrete-time models disappear when using continuous time (see Exercise 2.21). Moreover, continuous-time models have more flexibility in the analysis of dynamics and allow explicit-form solutions in a wider set of circumstances. These considerations motivate the detailed study of both the discrete- and continuous-time versions of the basic models in this book.

Let us start with a simple difference equation:

$$x(t+1) - x(t) = g(x(t)). \tag{2.30}$$

This equation states that between time t and $t+1$, the absolute growth in x is given by $g(x(t))$. Imagine that time is more finely divisible than that captured by our discrete indices, $t = 0, 1, \ldots$. In the limit, we can think of time as being as finely divisible as we would like, so that $t \in \mathbb{R}_+$. In that case, (2.30) gives us information about how the variable x changes between two discrete points in time, t and $t+1$. Between these time periods, we do not know how x evolves. However, if t and $t+1$ are not too far apart, the following approximation is reasonable:

$$x(t + \Delta t) - x(t) \simeq \Delta t \cdot g(x(t))$$

for any $\Delta t \in [0, 1]$. When $\Delta t = 0$, this equation is just an identity. When $\Delta t = 1$, it gives (2.30). In between it is a linear approximation. This approximation will be relatively accurate if the distance between t and $t+1$ is not very large, so that $g(x) \simeq g(x(t))$ for all $x \in [x(t), x(t+1)]$ (however, you should also convince yourself that this approximation could in fact be quite bad if the function g is highly nonlinear, in which case its behavior changes significantly between $x(t)$ and $x(t+1)$). Now divide both sides of this equation by Δt, and take limits to obtain

$$\lim_{\Delta t \to 0} \frac{x(t + \Delta t) - x(t)}{\Delta t} = \dot{x}(t) \simeq g(x(t)), \tag{2.31}$$

where, as throughout the book, I use the dot notation to denote time derivatives, $\dot{x}(t) \equiv dx(t)/dt$. Equation (2.31) is a differential equation representing the same dynamics as the difference equation (2.30) for the case in which the distance between t and $t+1$ is small.

2.4.2 The Fundamental Equation of the Solow Model in Continuous Time

We can now repeat all of the analysis so far using the continuous-time representation. Nothing has changed on the production side, so we continue to have (2.6) and (2.7) as the factor prices, but now these refer to instantaneous rental rates. For example, $w(t)$ is the flow of wages that workers receive at instant t. Savings are again given by

$$S(t) = sY(t),$$

while consumption is still given by (2.11).

Let us also introduce population growth into this model and assume that the labor force $L(t)$ grows proportionally, that is,

$$L(t) = \exp(nt)L(0). \tag{2.32}$$

The purpose of doing so is that in many of the classical analyses of economic growth, population growth plays an important role, so it is useful to see how it affects the equilibrium here. There is still no technological progress.

Recall that

$$k(t) \equiv \frac{K(t)}{L(t)},$$

which implies that

$$\frac{\dot{k}(t)}{k(t)} = \frac{\dot{K}(t)}{K(t)} - \frac{\dot{L}(t)}{L(t)},$$
$$= \frac{\dot{K}(t)}{K(t)} - n,$$

where I used the fact that, from (2.32), $\dot{L}(t)/L(t) = n$. From the limiting argument leading to equation (2.31) in the previous subsection, the law of motion of the capital stock is given by

$$\dot{K}(t) = sF(K(t), L(t), A(t)) - \delta K(t).$$

Using the definition of $k(t)$ as the capital-labor ratio and the constant returns to scale properties of the production function, the fundamental law of motion of the Solow model in continuous time is obtained as

$$\frac{\dot{k}(t)}{k(t)} = s\frac{f(k(t))}{k(t)} - (n + \delta), \tag{2.33}$$

where, following usual practice, I have transformed the left-hand side to the proportional change in the capital-labor ratio by dividing both sides by $k(t)$.[6]

Definition 2.5 *In the basic Solow model in continuous time with population growth at the rate n, no technological progress and an initial capital stock $K(0)$, an* equilibrium path *is given by paths (sequences) of capital stocks, labor, output levels, consumption levels, wages, and rental rates $[K(t), L(t), Y(t), C(t), w(t), R(t)]_{t=0}^{\infty}$ such that $L(t)$ satisfies (2.32), $k(t) \equiv K(t)/L(t)$ satisfies (2.33), $Y(t)$ is given by (2.1), $C(t)$ is given by (2.11), and $w(t)$ and $R(t)$ are given by (2.6) and (2.7), respectively.*

As before, a steady-state equilibrium involves $k(t)$ remaining constant at some level k^*.

It is easy to verify that the equilibrium differential equation (2.33) has a unique steady state at k^*, which is given by a slight modification of (2.18) to incorporate population growth:

$$\frac{f(k^*)}{k^*} = \frac{n + \delta}{s}. \tag{2.34}$$

In other words, going from discrete to continuous time has not changed any of the basic economic features of the model. Thus the steady state can again be plotted in a diagram similar to Figure 2.1 except that it now also incorporates population growth. This is done in Figure 2.8, which also highlights that the logic of the steady state is the same with population growth as it was without population growth. The amount of investment, $sf(k)$, is used to replenish the capital-labor ratio, but now there are two reasons for replenishments. The capital stock depreciates exponentially at the flow rate δ. In addition, the capital stock must also increase as

6. Throughout I adopt the notation $[x(t)]_{t=0}^{\infty}$ to denote the continuous-time path of variable $x(t)$. An alternative notation often used in the literature is $(x(t); t \geq 0)$. I prefer the former both because it is slightly more compact and also because it is more similar to the discrete-time notation for the time path of a variable, $\{x(t)\}_{t=0}^{\infty}$. When referring to $[x(t)]_{t=0}^{\infty}$, I use the terms "path," "sequence," and "function (of time t)" interchangeably.

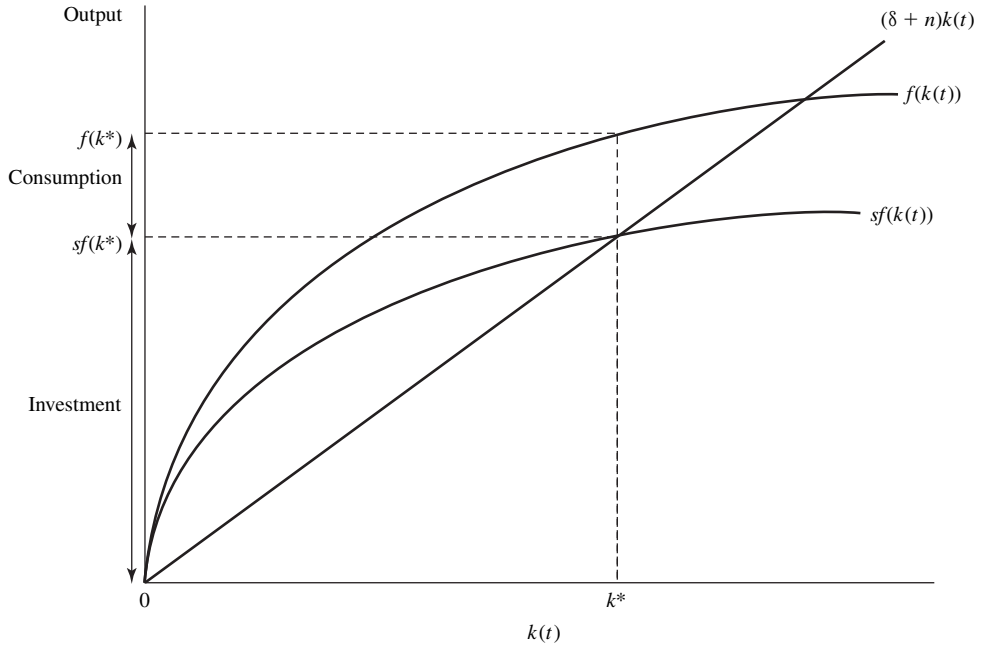

FIGURE 2.8 Investment and consumption in the steady-state equilibrium with population growth.

population grows to maintain the capital-labor ratio at a constant level. The amount of capital that needs to be replenished is therefore $(n + \delta)k$.

Proposition 2.7 *Consider the basic Solow growth model in continuous time and suppose that Assumptions 1 and 2 hold. Then there exists a unique steady-state equilibrium where the capital-labor ratio is equal to* $k^* \in (0, \infty)$ *and satisfies (2.34), per capita output is given by*

$$y^* = f(k^*),$$

and per capita consumption is given by

$$c^* = (1 - s) f(k^*).$$

Proof. See Exercise 2.5. ■

Moreover, again defining $f(k) = A\tilde{f}(k)$, the following proposition holds.

Proposition 2.8 *Suppose Assumptions 1 and 2 hold and* $f(k) = A\tilde{f}(k)$. *Denote the steady-state equilibrium level of the capital-labor ratio by* $k^*(A, s, \delta, n)$ *and the steady-state level of output by* $y^*(A, s, \delta, n)$ *when the underlying parameters are given by A, s, δ, and n. Then we have*

$$\frac{\partial k^*(A,s,\delta,n)}{\partial A} > 0, \quad \frac{\partial k^*(A,s,\delta,n)}{\partial s} > 0, \quad \frac{\partial k^*(A,s,\delta,n)}{\partial \delta} < 0, \quad \text{and} \quad \frac{\partial k^*(A,s,\delta,n)}{\partial n} < 0;$$
$$\frac{\partial y^*(A,s,\delta,n)}{\partial A} > 0, \quad \frac{\partial y^*(A,s,\delta,n)}{\partial s} > 0, \quad \frac{\partial y^*(A,s,\delta,n)}{\partial \delta} < 0, \quad \text{and} \quad \frac{\partial y^*(A,s,\delta,n)}{\partial n} < 0.$$

Proof. See Exercise 2.6. ■

The new result relative to the earlier comparative static proposition (Proposition 2.3) is that now a higher population growth rate, n, also reduces the capital-labor ratio and output per