1 INTRODUCTION: THE FACTS OF ECONOMIC GROWTH

> The errors which arise from the absence of facts are far more numerous and more durable than those which result from unsound reasoning respecting true data.
>
> — CHARLES BABBAGE, quoted in Rosenberg (1994), p. 27.

> It is quite wrong to try founding a theory on observable magnitudes alone.... It is the theory which decides what we can observe.
>
> — ALBERT EINSTEIN, quoted in Heisenberg (1971), p. 63.

Speaking at the annual meeting of the American Economic Association in 1989, the renowned economic historian David S. Landes chose as the title of his address the fundamental question of economic growth and development: "Why Are We So Rich and They So Poor?"[1] This age-old question has preoccupied economists for centuries. It so fascinated the classical economists that it was stamped on the cover of Adam Smith's famous treatise *An Inquiry into the Nature and Causes of the Wealth of Nations*. And it was the mistaken forecast of Thomas Malthus in the early nineteenth century concerning the future prospects for economic growth that earned the discipline its most recognized epithet, the "dismal science."

[1] See Landes (1990).

The modern examination of this question by macroeconomists dates to the 1950s and the publication of two famous papers by Robert Solow of the Massachusetts Institute of Technology. Solow's theories helped to clarify the role of the accumulation of physical capital and emphasized the importance of technological progress as the ultimate driving force behind sustained economic growth. During the 1960s and to a lesser extent the 1970s, work on economic growth flourished.[2] For methodological reasons, however, important aspects of the theoretical exploration of technological change were postponed.[3]

In the early 1980s, work at the University of Chicago by Paul Romer and Robert Lucas re-ignited the interest of macroeconomists in economic growth, emphasizing the economics of "ideas" and of human capital. Taking advantage of new developments in the theory of imperfect competition, Romer introduced the economics of technology to macroeconomists. Following these theoretical advances, empirical work by a number of economists, such as Robert Barro of Harvard University, quantified and tested the theories of growth. Both theoretical and empirical work has since continued with enormous professional interest.

The purpose of this book is to explain and explore the modern theories of economic growth. This exploration is an exciting journey, in which we encounter several ideas that have already earned Nobel Prizes and several more with Nobel potential. The book attempts to make this cutting-edge research accessible to readers with only basic training in economics and calculus.[4]

The approach of this book is similar to the approach scientists take in studying astronomy and cosmology. Like economists, astronomers are unable to perform the controlled experiments that are the hallmark of chemistry and physics. Astronomy proceeds instead through an interplay between observation and theory. There is observation: planets,

[2]A far from exhaustive list of contributors includes Moses Abramovitz, Kenneth Arrow, David Cass, Tjalling Koopmans, Simon Kuznets, Richard Nelson, William Nordhaus, Edmund Phelps, Karl Shell, Eytan Sheshinski, Trevor Swan, Hirofumi Uzawa, and Carl von Weizsacker.

[3]Romer (1994) provides a nice discussion of this point and of the history of research on economic growth.

[4]The reader with advanced training is referred also to the excellent presentations in Barro and Sala-i-Martin (1998) and Aghion and Howitt (1998).

stars, and galaxies are laid out across the universe in a particular way. Galaxies are moving apart, and the universe appears to be sparsely populated with occasional "lumps" of matter. And there is theory: the theory of the Big Bang, for example, provides a coherent explanation for these observations.

This same interplay between observation and theory is used to organize this book. This first chapter will outline the broad empirical regularities associated with growth and development. How rich are the rich countries, how poor are the poor? How fast do rich and poor countries grow? The remainder of the book consists of theories to explain these observations. In the limited pages we have before us, we will not spend much time on the experiences of individual countries, although these experiences are very important. Instead, the goal is to provide a general economic framework to help us understand the process of growth and development.

A critical difference between astronomy and economics, of course, is that the economic "universe" can potentially be re-created by economic policy. Unlike the watchmaker who builds a watch and then leaves it to run forever, economic policy makers constantly shape the course of growth and development. A prerequisite to better policies is a better understanding of economic growth.

1.1 THE DATA OF GROWTH AND DEVELOPMENT

The world consists of economies of all shapes and sizes. Some countries are very rich, and some are very poor. Some economies are growing rapidly, and some are not growing at all. Finally, a large number of economies — most, in fact — lie between these extremes. In thinking about economic growth and development, it is helpful to begin by considering the extreme cases: the rich, the poor, and the countries that are moving rapidly in between. The remainder of this chapter lays out the empirical evidence — the "facts" — associated with these categories. The key questions of growth and development then almost naturally ask themselves.

Table 1.1 displays some basic data on growth and development for seventeen countries. We will focus our discussion of the data on measures of per capita income instead of reporting data such as life

TABLE 1.1 STATISTICS ON GROWTH AND DEVELOPMENT

	GDP per capita, 1997	GDP per worker, 1997	Labor force participation rate, 1997	Average annual growth rate, 1960–97	Years to double
"Rich" countries					
U.S.A.	$20,049	$40,834	0.49	1.4	50
Japan	16,003	25,264	0.63	4.4	16
France	14,650	31,986	0.46	2.3	30
U.K.	14,472	29,295	0.49	1.9	37
Spain	10,685	29,396	0.36	3.5	20
"Poor" countries					
China	2,387	3,946	0.60	3.5	20
India	1,624	4,156	0.39	2.3	30
Zimbabwe	1,242	2,561	0.49	0.4	192
Uganda	697	1,437	0.49	0.5	146
"Growth miracles"					
Hong Kong	18,811	28,918	0.65	5.2	13
Singapore	17,559	36,541	0.48	5.4	13
Taiwan	11,729	26,779	0.44	5.6	12
South Korea	10,131	24,325	0.42	5.9	12
"Growth disasters"					
Venezuela	6,760	19,455	0.35	−0.1	−517
Madagascar	577	1,334	0.43	−1.5	−46
Mali	535	1,115	0.48	−0.8	−85
Chad	392	1,128	0.35	−1.4	−48

SOURCE: Author's calculations using Penn World Tables Mark 5.6, an update of Summers and Heston (1991), and the World Bank's Global Development Network Growth Database, assembled by William Easterly and Hairong Yu.

Notes: The GDP data are in 1985 dollars. The growth rate is the average annual change in the log of GDP per worker. A negative number in the "Years to double" column indicates "years to halve."

expectancy, infant mortality, or other measures of quality of life. The main reason for this focus is that the theories we develop in subsequent chapters will be couched in terms of per capita income. Furthermore, per capita income is a useful "summary statistic" of the level of economic development in the sense that it is highly correlated with other measures of quality of life.[5]

We will interpret Table 1.1 in the context of some "facts," beginning with the first:[6]

FACT #1 **There is enormous variation in per capita income across economies. The poorest countries have per capita incomes that are less than 5 percent of per capita incomes in the richest countries.**

The first section of Table 1.1 reports real per capita gross domestic product (GDP) in 1997, together with some other data, for the United States and several other "rich" countries. The United States was the richest country in the world in 1997, with a per capita GDP of $20,049 (in 1985 dollars), and it was the richest by a substantial amount. Japan, for example had a per capita GDP of about $16,000.

These numbers may at first seem slightly surprising. One sometimes reads in newspapers that the United States has fallen behind countries like Japan or Germany in terms of per capita income. Such newspaper accounts can be misleading, however, because market exchange rates are typically used in the comparison. U.S. GDP is measured in dollars, whereas Japanese GDP is measured in yen. How do we convert the Japanese yen to dollars in order to make a comparison? One way is to use prevailing exchange rates. For example, in January 1997, the yen-to-dollar exchange rate was around 120 yen per dollar. However, exchange rates can be extremely volatile. Just a little over one year earlier, the rate was only 100 yen per dollar. Which of these exchange rates is "right"?

[5] See, for example, the World Bank's *World Development Report, 1991* (New York: Oxford University Press, 1991).

[6] Many of these facts have been discussed elsewhere. See especially Lucas (1988) and Romer (1989).

Obviously, it matters a great deal which one we use: at 100 yen per dollar, Japan will seem 20 percent richer than at 120 yen per dollar.

Instead of relying on prevailing exchange rates to make international comparisons of GDP, economists attempt to measure the actual value of a currency in terms of its ability to purchase similar products. The resulting conversion factor is sometimes called a purchasing power parity–adjusted exchange rate. For example, the *Economist* magazine produces a yearly report of purchasing power parity (PPP) exchange rates based on the price of a McDonald's Big Mac hamburger. If a Big Mac costs 2 dollars in the United States and 300 yen in Japan, then the PPP exchange rate based on the Big Mac is 150 yen per dollar. By extending this method to a number of different goods, economists construct a PPP exchange rate that can be applied to GDP. Such calculations suggest that 150 yen per dollar is a much better number than the prevailing exchange rates of 100 or 120 yen per dollar.[7]

The second column of Table 1.1 reports a related measure, real GDP per worker in 1997. The difference between the two columns lies in the denominator: the first column divides total GDP by a country's entire population, while the second column divides GDP by only the labor force. The third column reports the 1997 labor force participation rate — the ratio of the labor force to the population — to show the relationship between the first two columns. Notice that while Japan had a higher per capita GDP than France in 1997, the comparison for GDP per worker is reversed. The labor force participation rate is much higher in Japan than in the other industrialized countries.

Which column should we use in comparing levels of development? The answer depends on what question is being asked. Perhaps per capita GDP is a more general measure of welfare in that it tells us how much output per person is available to be consumed, invested, or put to some other use. On the other hand, GDP per worker tells us more about the productivity of the labor force. In this sense, the first statistic can be thought of as a welfare measure, while the second is a productivity measure. This seems to be a reasonable way to interpret these statistics, but one can also make the case for using GDP per worker as a welfare measure. Persons not officially counted as being in the labor force may be engaged in "home production" or may work in the underground

[7] *Economist*, April 19, 1995, p. 74.

economy. Neither of these activities is included in GDP, and in this case measured output divided by measured labor input may prove more accurate for making welfare comparisons. In this book, we will often use the phrase "per capita income" as a generic welfare measure, even when speaking of GDP per worker, if the context is clear. Whatever measure we use, though, Table 1.1 tells us one of the first key things about economic development: the more "effort" an economy puts into producing output, the more output there is to go around. "Effort" in this context corresponds to the labor force participation rate.

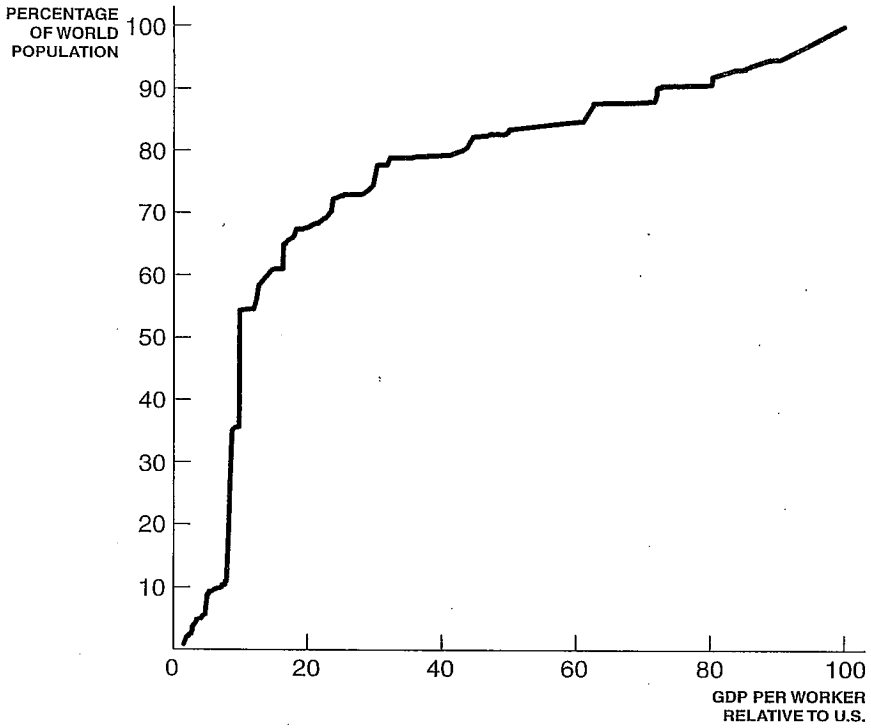
The second section of Table 1.1 documents the relative and even absolute poverty of some of the world's poorest economies. India and Zimbabwe had per capita GDPs around $1,500 in 1997, less than 10 percent of that in the United States. A number of economies in sub-Saharan Africa are even poorer: per capita income in the United States is more than 40 times higher than income in Ethiopia.

To place these numbers in perspective, consider some other statistics. The typical worker in Ethiopia or Uganda must work a month and a half to earn what the typical worker in the United States earns in a day. Life expectancy in Ethiopia is only two-thirds that in the United States, and infant mortality is more than 20 times higher. Approximately 40 percent of GDP is spent on food in Ethiopia, compared to about 7 percent in the United States.

What fraction of the world's population lives with this kind of poverty? Figure 1.1 answers this question by plotting the distribution of the world's population in terms of GDP per worker. In 1995, more than half of the world's population lived in countries with less than 10 percent of U.S. GDP per worker. The bulk of this population lives in only two countries: China, with nearly one-quarter of the world's population, and India, with one-sixth of the world's population. Together, these two countries account for more than 40 percent of the world's population. In contrast, the 39 countries that make up sub-Saharan Africa constitute about 10 percent of the world's population.

Figure 1.2 shows how this distribution has changed since 1960. Overall, the distribution has equalized as the share of the world's population living in countries whose GDP per worker is less than 30 percent of that in the United States has fallen. Of the poorest countries, both China and India have seen substantial growth in GDP per worker, even relative to the United States. China's relative income rose from 4 percent

FIGURE 1.1 CUMULATIVE DISTRIBUTION OF WORLD POPULATION BY GDP PER WORKER, 1995



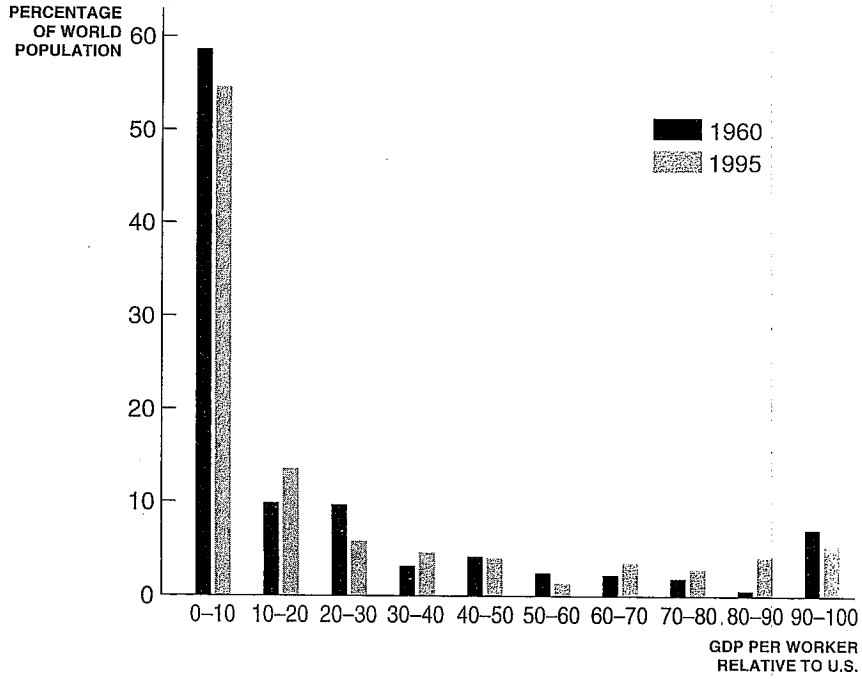
SOURCE: Penn World Tables Mark 5.6, Summers and Heston (1991), updated using Easterly and Yu (2000).

Note: A point (x, y) in the figure indicates that the fraction of the world's population living in countries with a relative GDP per worker less than x is equal to y. 136 countries are included.

of U.S. GDP per worker in 1960 to 9 percent in 1995, and India's relative income rose from 7 percent of U.S. GDP per worker to 10 percent over the same period.

The third section of Table 1.1 reports data for several countries that are moving from the second group to the first. These four so-called newly industrializing countries (NICs) are Hong Kong, Singapore, Taiwan, and South Korea. Interestingly, by 1997 Hong Kong had a per

FIGURE 1.2 WORLD POPULATION BY GDP PER WORKER, 1960 AND 1995



SOURCE: Penn World Tables Mark 5.6, Summers and Heston (1991), updated using Easterly and Yu (2000).

Note: The sample size has been reduced to 114 countries in order to incorporate the 1960 data.

capita GDP of $18,811, higher than all of the industrialized countries in the table except for the United States. This per capita GDP was almost twice that of South Korea. However, as with Japan, Hong Kong's high per capita GDP is driven to a large extent by its high labor force participation rate. In terms of GDP per worker, Hong Kong comes in below the other industrialized economies. Singapore, on the other hand, has a GDP per worker of $36,541, one of the highest in the world.

An important characteristic of these NICs is their extremely rapid rates of growth, and this leads to our next fact:

FACT #2 Rates of economic growth vary substantially across countries.

The last two columns of Table 1.1 characterize economic growth. The fourth column reports the average annual change in the (natural) log of GDP per worker from 1960 to 1997.[8] Growth in GDP per worker in the United States averaged only 1.4 percent per year from 1960 to 1997. France, the United Kingdom, and Spain grew a bit more rapidly, while Japan grew at a remarkable rate of 4.4 percent. The NICs exceeded even Japan's astounding rate of increase, truly exemplifying what is meant by the term "growth miracle." The poorest countries of the world exhibited varied growth performance. China and India, for example, grew substantially faster than the United States from 1960 to 1997, but their growth rates were well below those of the NICs. Other developing countries such as Zimbabwe and Uganda experienced little or no growth over the period. Finally, growth rates in a number of countries were negative from 1960 to 1997, earning these countries the label "growth disasters." Real incomes actually declined in countries such as Venezuela, Madagascar, and Chad, as shown in the last panel of Table 1.1.

A useful way to interpret these growth rates was provided by Robert E. Lucas, Jr., in a paper titled "On the Mechanics of Economic Development" (1988). A convenient rule of thumb used by Lucas is that a country growing at g percent per year will double its per capita income every $70/g$ years.[9] According to this rule, U.S. GDP per worker will

[8]See Appendix A for a discussion of how this concept of growth relates to percentage changes.

[9]Let $y(t)$ be per capita income at time t and let y_0 be some initial value of per capita income. Then $y(t) = y_0 e^{gt}$. The time it takes per capita income to double is given by the time t^* at which $y(t) = 2y_0$. Therefore,

$$2y_0 = y_0 e^{gt^*}$$

$$\Longrightarrow t^* = \frac{\log 2}{g}.$$

The rule of thumb is established by noting that $\log 2 \approx .7$. See Appendix A for further discussion.

double approximately every 50 years, while Korean GDP per worker will double approximately every 12 years. In other words, if these growth rates persisted for two generations, the average American would be two or three times as rich as his or her grandparents. The average citizen of Taiwan, Hong Kong, or South Korea would be *twenty* times as rich as his or her grandparents. Over moderate periods of time, small differences in growth rates can lead to enormous differences in per capita incomes.

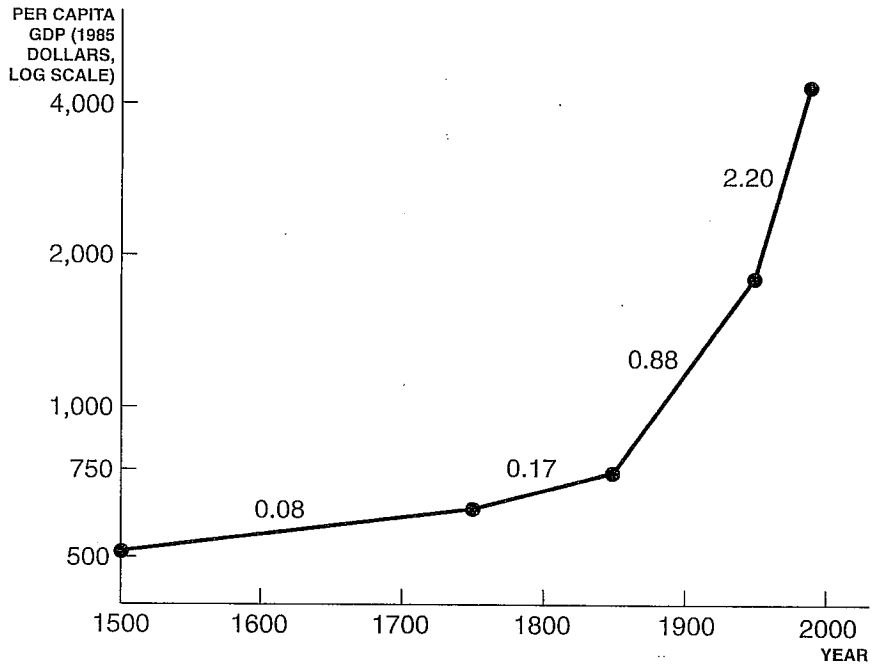
FACT #3 Growth rates are not generally constant over time. For the world as a whole, growth rates were close to zero over most of history but have increased sharply in the twentieth century. For individual countries, growth rates also change over time.

The rapid growth rates observed in East Asia—and even the more modest growth rates of about 2 percent per year observed throughout the industrialized world—are blindingly fast when placed in a broad historical context. Figure 1.3 illustrates this point by plotting a measure of world GDP per capita over the past five centuries. Notice that because the graph is plotted on a log scale, the slope of each line segment reflects the rate of growth: the rising slope over time indicates a rise in the world's economic growth rate.

Between 1950 and 1990, world per capita GDP grew at a rate of 2.2 percent per year. Between 1850 and 1950, however, the growth rate was only 0.88 percent, and before 1850 the growth rate was less than 0.2 percent per year. Angus Maddison (1995) goes so far as to suggest that during the millennium between 500 and 1500, growth was essentially zero. Sustained economic growth at rates of 2 percent per year is just as much a modern invention as is electricity or the transistor.

As a result of this growth, the world is substantially richer today than it has ever been before. A rough guess is that per capita GDP for the world as a whole in 1500 was $500 per person. Today, world per capita GDP is nearly ten times higher.

FIGURE 1.3 **WORLD PER CAPITA GDP AND GROWTH RATES, 1500–1990**



SOURCE: Computed from Lucas (1998) and Maddison (1995).
Note: The numbers above each line segment are average annual growth rates.

As a rough check on these numbers, consider the following exercise. Suppose we guess that the world, or even a particular country, has grown at a rate of 2 percent per year forever. This means that per capita income must have been doubling every 35 years. Over the last 250 years, income would have grown by a factor of about 2^7, or 128. In this case, an economy with a per capita GDP of $20,000 today would have had a per capita GDP of just over $150 in 1750, measured at today's prices—less than half the per capita GDP of the poorest countries in the world today. It is virtually impossible to live on 50 cents per day, and so we know that a growth rate of 2 percent per year could not have been sustained even for 250 years.

For individual countries, growth rates also change over time, as can be seen in a few interesting examples. India's average growth rate from 1960 to 1997 was 2.3 percent per year. From 1960 to 1980, however, its growth rate was only 1.3 percent per year; between 1980 and 1997 growth accelerated to 3.5 percent per year. Singapore did not experience particularly rapid growth until after the 1950s. The island country of Mauritius exhibited a strong *decline* in GDP per worker of 1.3 percent per year in the two decades following 1950. From 1970 to 1997, however, Mauritius grew at 3.6 percent per year. Finally, economic reforms in China have had a substantial impact on growth and on the economic well-being of one-quarter of the world's population. Between 1960 and 1978, GDP per worker grew at an annual rate of 1.9 percent in China. Since 1979, however, growth has averaged 5.0 percent per year.

The substantial variation in growth rates both across and within countries leads to an important corollary of Facts 2 and 3. It is so important that we will call it a fact itself:

FACT #4 **A country's relative position in the world distribution of per capita incomes is not immutable. Countries can move from being "poor" to being "rich," and vice-versa.**[10]

1.2 OTHER "STYLIZED FACTS"

Facts 1 through 4 apply broadly to the countries of the world. The next fact describes some features of the U.S. economy. These features turn out to be extremely important, as we will see in Chapter 2. They are general characteristics of most economies "in the long run."

[10] A classic example of the latter is Argentina. At the end of the nineteenth century, Argentina was one of the richest countries in the world. With a tremendous natural resource base and a rapidly developing infrastructure, it attracted foreign investment and immigration on a large scale. By 1990, however, Argentina's per capita income was only about one-third of per capita income in the United States. Carlos Diaz-Alejandro (1970) provides a classic discussion of the economic history of Argentina.

FACT #5 **In the United States over the last century,**

1. **the real rate of return to capital, r, shows no trend upward or downward;**
2. **the shares of income devoted to capital, rK/Y, and labor, wL/Y, show no trend; and**
3. **the average growth rate of output per person has been positive and relatively constant over time — i.e., the United States exhibits steady, sustained per capita income growth.**

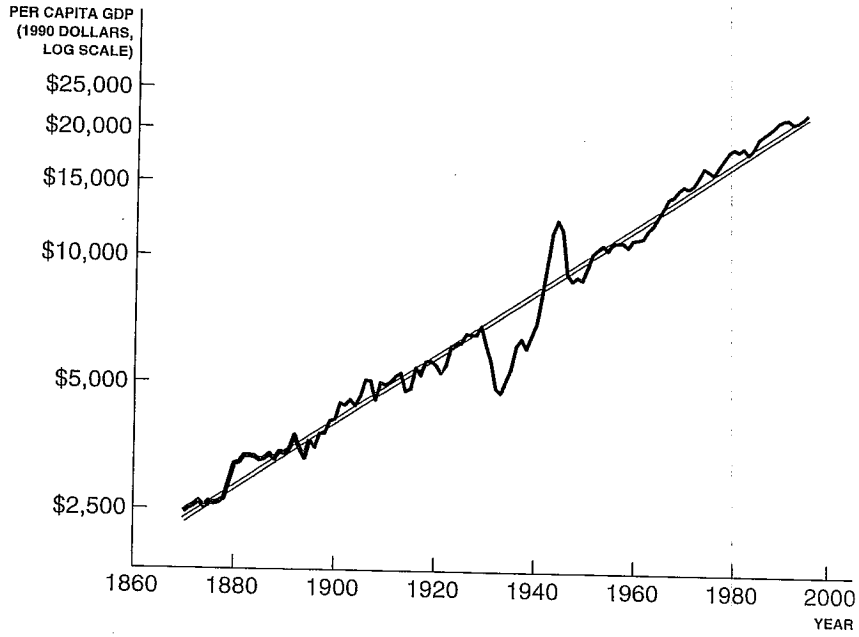
This stylized fact, really a collection of facts, is drawn largely from a lecture given by Nicholas Kaldor at a 1958 conference on capital accumulation (Kaldor 1961). Kaldor, following the advice of Charles Babbage, began the lecture by claiming that the economic theorist should begin with a summary of the "stylized" facts a theory was supposed to explain.

Kaldor's first fact — that the rate of return to capital is roughly constant — is best seen by noting that the real interest rate on government debt in the U.S. economy shows no trend. Granted, we do not observe real interest rates, but one can take the nominal interest rate and subtract off either the expected or the actual rate of inflation to make this observation.

The second fact concerns payments to the factors of production, which we can group into capital and labor. For the United States, one can calculate labor's share of GDP by looking at wage and salary payments and compensation for the self-employed as a share of GDP.[11] These calculations reveal that the labor share has been relatively constant over time, at a value of around 0.7. If we are focusing on a model with two factors, and if we assume that there are no economic profits in the model, then the capital share is simply 1 minus the labor share, or 0.3. These first two facts imply that the capital-output ratio, K/Y, is roughly constant in the United States.

[11]These data are reported in the National Income and Product Accounts. See, for example, the Council of Economic Advisors (1997).

FIGURE 1.4 **REAL PER CAPITA GDP IN THE UNITED STATES, 1870–1994**



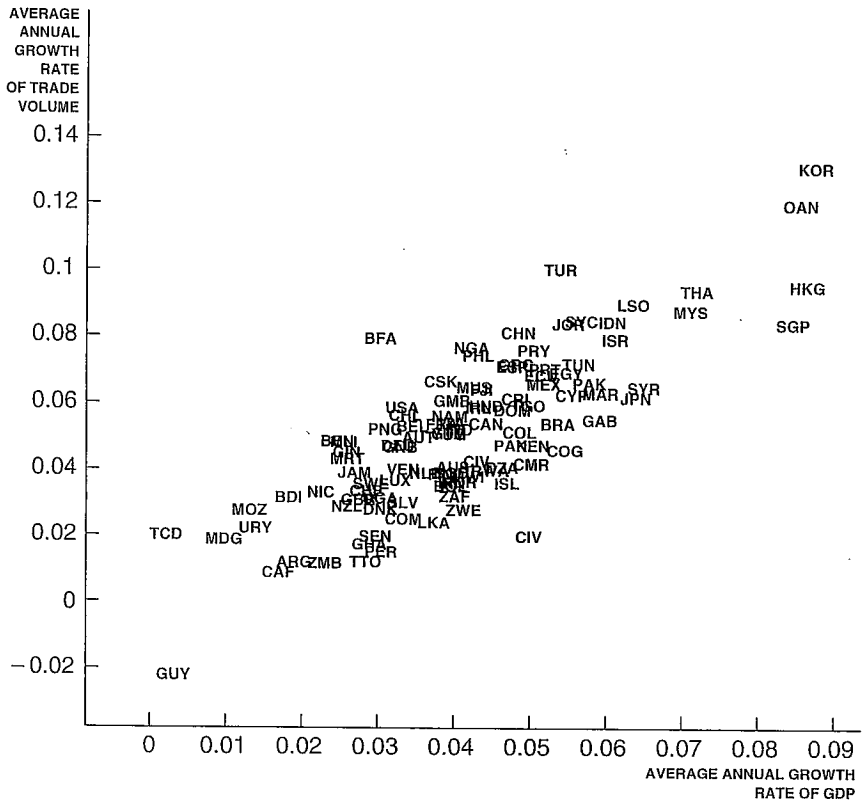
SOURCE: Maddison (1995) and author's calculations.

The third fact is a slight reinterpretation of one of Kaldor's stylized facts, illustrated in Figure 1.4. The figure plots per capita GDP (on a log scale) for the United States from 1870 until 1994. The trend line in the figure rises at a rate of 1.8 percent per year, and the relative constancy of the growth rate can be seen by noting that apart from the ups and downs of business cycles, this constant growth rate path "fits" the data very well.

FACT #6 **Growth in output and growth in the volume of international trade are closely related.**

Figure 1.5 documents the close relationship between the growth in a country's output (GDP) and growth in its volume of trade. Here, the

FIGURE 1.5 **GROWTH IN TRADE AND GDP, 1960–90**



SOURCE: Penn World Tables Mark 5.6, Summers and Heston (1991).

volume of trade is defined as the sum of exports and imports, but a similar figure could be produced with either component of trade. Notice that for many countries, trade volume has grown faster than GDP; the share of exports and imports in GDP has generally increased around the world since 1960.[12]

[12]On this point, it is interesting to note that the world economy was very open to international trade prior to World War I. Jeffrey Sachs and Andrew Warner (1995) argue that much of the trade liberalization since World War II, at least until the 1980s, simply re-establishes the global nature of markets that prevailed in 1900.

The relationship between trade and economic performance is complicated. Some economies, such as those of Hong Kong, Singapore, and Luxembourg, have flourished as regional "trade centers." The trade intensity ratio—the sum of exports and imports divided by GDP—for these economies *exceeds* 150 percent. How is this possible? These economies import unfinished products, add value by completing the production process, and then export the result. GDP, of course, is generated only in the second step. A substantial component of the strong growth performance turned in by these economies is associated with an increase in trade intensity.

On the other hand, trade intensity in Japan actually fell from around 21 percent in 1960 to around 18 percent in 1992 despite rapid per capita growth. And nearly all of the countries in sub-Saharan Africa have trade intensities higher than Japan's. A number of these countries also saw trade intensity increase from 1960 to 1990 while economic growth faltered.

FACT #7 **Both skilled and unskilled workers tend to migrate from poor to rich countries or regions.**

Robert Lucas emphasized this stylized fact in his aforementioned article. Evidence for the fact can be seen in the presence of in-migration restrictions in rich countries. It is an important observation because these movements of labor, which presumably are often very costly, tell us something about real wages. The returns to both skilled and unskilled labor must be *higher* in high-income regions than in low-income regions. Otherwise, labor would not be willing to pay the high costs of migration. In terms of skilled labor, this raises an interesting puzzle. Presumably, skilled labor is scarce in developing economies, and simple theories predict that factor returns are highest where factors are scarce. Why, then, doesn't skilled labor migrate from the United States to Zaire?

THE REMAINDER OF THIS BOOK

Three central questions of economic growth and development are examined in the remainder of this book.

The first question is the one asked at the beginning of this chapter: Why are we so rich and they so poor? It is a question about *levels* of development and the world distribution of per capita incomes. This topic is explored in Chapters 2 and 3 and then is revisited in Chapter 7.

The second question is, What is the engine of economic growth? How is it that economies experience sustained growth in output per worker over the course of a century or more? Why is it that the United States has grown at 1.8 percent per year since 1870? The answer to these questions is *technological progress*. Understanding why technological progress occurs and how a country such as the United States can exhibit sustained growth is the subject of Chapters 4 and 5.

The final question concerns *growth miracles*. How is it that economies such as Japan's after World War II and those of Hong Kong, Singapore, and South Korea more recently are able to transform rapidly from "poor" to "rich?" Such Cinderella-like transformation gets at the heart of economic growth and development. Chapters 6 and 7 present one theory that integrates the models of the earlier chapters.

The next two chapters depart from the cumulative flow of the book to explore new directions. Chapter 8 discusses influential alternative theories of economic growth. Chapter 9 examines the potentially important interactions between natural resources and the sustainability of growth. Chapter 10 offers some conclusions.

Three appendices complete this book. Appendix A reviews the mathematics needed throughout the book.[13] Appendix B lists a number of very readable articles and books related to economic growth that make excellent supplementary reading. And Appendix C presents a collection of the data analyzed throughout the book. The country codes used in figures such as Figure 1.5 are also translated there.

The facts we have examined in this chapter indicate that it is not simply out of intellectual curiosity that we ask these questions. The

[13]Readers with a limited exposure to calculus, differential equations, and the mathematics of growth are encouraged to read Appendix A before continuing with the next chapter.

answers hold the key to unlocking widespread rapid economic growth. Indeed, the recent experience of East Asia suggests that such growth has the power to transform standards of living over the course of a single generation. Surveying this evidence in the 1985 Marshall Lecture at Cambridge University, Robert E. Lucas, Jr., expressed the sentiment that fueled research on economic growth for the next decade:

> I do not see how one can look at figures like these without seeing them as representing *possibilities*. Is there some action a government of India could take that would lead the Indian economy to grow like Indonesia's or Egypt's? If so, *what* exactly? If not, what is it about the "nature of India" that makes it so? The consequences for human welfare involved in questions like these are simply staggering: Once one starts to think about them, it is hard to think about anything else (Lucas 1988, p. 5).

2 THE SOLOW MODEL

> All theory depends on assumptions which are not quite true. That is what makes it theory. The art of successful theorizing is to make the inevitable simplifying assumptions in such a way that the final results are not very sensitive. —ROBERT SOLOW (1956), P. 65.

In 1956, Robert Solow published a seminal paper on economic growth and development titled "A Contribution to the Theory of Economic Growth." For this work and for his subsequent contributions to our understanding of economic growth, Solow was awarded the Nobel Prize in economics in 1987. In this chapter, we develop the model proposed by Solow and explore its ability to explain the stylized facts of growth and development discussed in Chapter 1. As we will see, this model provides an important cornerstone for understanding why some countries flourish while others are impoverished.

Following the advice of Solow in the quotation above, we will make several assumptions that may seem to be heroic. Nevertheless, we hope that these are simplifying assumptions in that, for the purposes at hand, they do not terribly distort the picture of the world we create. For example, the world we consider in this chapter will consist of countries that produce and consume only a single, homogeneous good (*output*). Conceptually, as well as for testing the model using empirical data, it is convenient to think of this output as units of a country's gross domestic product, or GDP. One implication of this simplifying assumption is

that there is no international trade in the model because there is only a single good: I'll give you a 1941 Joe DiMaggio autograph in exchange for . . . your 1941 Joe DiMaggio autograph? Another assumption of the model is that technology is *exogenous*—that is, the technology available to firms in this simple world is unaffected by the actions of the firms, including research and development (R&D). These are assumptions that we will relax later on, but for the moment, and for Solow, they serve well. Much progress in economics has been made by creating a very simple world and then seeing how it behaves and misbehaves.

Before presenting the Solow model, it is worth stepping back to consider exactly what a model is and what it is for. In modern economics, a model is a mathematical representation of some aspect of the economy. It is easiest to think of models as toy economies populated by robots. We specify exactly how the robots behave, which is typically to maximize their own utility. We also specify the constraints the robots face in seeking to maximize their utility. For example, the robots that populate our economy may want to consume as much output as possible, but they are limited in how much output they can produce by the techniques at their disposal. The best models are often very simple but convey enormous insight into how the world works. Consider the supply and demand framework in microeconomics. This basic tool is remarkably effective at predicting how the prices and quantities of goods as diverse as health care, computers, and nuclear weapons will respond to changes in the economic environment.

With this understanding of how and why economists develop models, we pause to highlight one of the important assumptions we will make until the final chapters of this book. Instead of writing down utility functions that the robots in our economy maximize, we will summarize the results of utility maximization with elementary rules that the robots obey. For example, a common problem in economics is for an individual to decide how much to consume today and how much to save for consumption in the future. Another is for individuals to decide how much time to spend going to school to accumulate skills and how much time to spend working in the labor market. Instead of writing these problems down formally, we will assume that individuals save a constant fraction of their income and spend a constant fraction of their time accumulating skills. These are extremely useful simplifications; without them, the models are difficult to solve without more advanced

mathematical techniques. For many purposes, these are fine assumptions to make in our first pass at understanding economic growth. Rest assured, however, that we will relax these assumptions in Chapter 7.

2.1 THE BASIC SOLOW MODEL

The Solow model is built around two equations, a production function and a capital accumulation equation. The production function describes how inputs such as bulldozers, semiconductors, engineers, and steelworkers combine to produce output. To simplify the model, we group these inputs into two categories, capital, K, and labor, L, and denote output as Y. The *production function* is assumed to have the Cobb-Douglas form and is given by

$$Y = F(K, L) = K^{\alpha}L^{1-\alpha}, \tag{2.1}$$

where α is some number between 0 and 1.[1] Notice that this production function exhibits *constant returns to scale:* if all of the inputs are doubled, output will exactly double.[2]

Firms in this economy pay workers a wage, w, for each unit of labor and pay r in order to rent a unit of capital for one period. We assume there are a large number of firms in the economy so that perfect competition prevails and the firms are price-takers.[3] Normalizing the price of output in our economy to unity, profit-maximizing firms solve the following problem:

$$\max_{K,L} F(K, L) - rK - wL.$$

According to the first-order conditions for this problem, firms will hire labor until the marginal product of labor is equal to the wage and will rent capital until the marginal product of capital is equal to the rental price:

$$w = \frac{\partial F}{\partial L} = (1 - \alpha)\frac{Y}{L},$$
$$r = \frac{\partial F}{\partial K} = \alpha\frac{Y}{K}.$$

Notice that $wL + rK = Y$. That is, payments to the inputs ("factor payments") completely exhaust the value of output produced so that there are no economic profits to be earned. This important result is a general property of production functions with constant returns to scale. Notice also that the share of output paid to labor is $wL/Y = 1 - \alpha$ and the share paid to capital is $rK/Y = \alpha$. These factor shares are therefore constant over time, consistent with Fact 5 from Chapter 1.

Recall from Chapter 1 that the stylized facts we are typically interested in explaining involve output per worker or per capita output. With this interest in mind, we can rewrite the production function in equation (2.1) in terms of output per worker, $y \equiv Y/L$, and capital per worker, $k \equiv K/L$:

$$y = k^{\alpha}. \tag{2.2}$$

This production function is graphed in Figure 2.1. With more capital per worker, firms produce more output per worker. However, there are diminishing returns to capital per worker: each additional unit of capital we give to a single worker increases the output of that worker by less and less.

The second key equation of the Solow model is an equation that describes how capital accumulates. The capital accumulation equation is given by

$$\dot{K} = sY - dK. \tag{2.3}$$

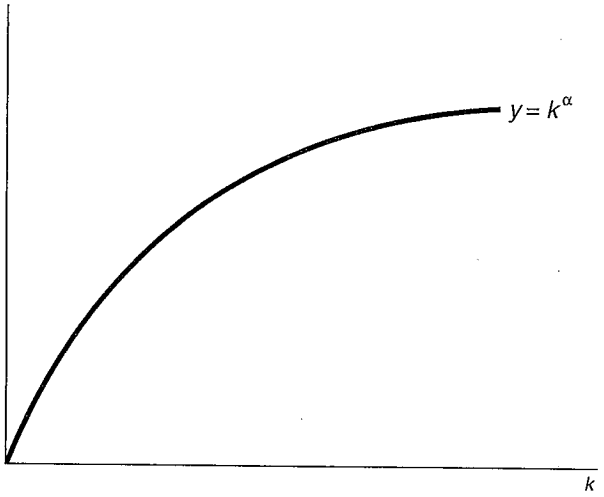
This kind of equation will be used throughout this book and is very important, so let's pause a moment to explain carefully what this equation says. According to this equation, the change in the capital stock, $\dot{K}$, is equal to the amount of gross investment, sY, less the amount of depreciation that occurs during the production process, dK. We'll now discuss these three terms in more detail.

[1] Charles Cobb and Paul Douglas (1928) proposed this functional form in their analysis of U.S. manufacturing. Interestingly, they argued that this production function, with a value for α of 1/4, fit the data very well without allowing for technological progress.

[2] Recall that if $F(aK, aL) = aY$ for any number $a > 1$, then we say that the production function exhibits constant returns to scale. If $F(aK, aL) > aY$, then the production function exhibits *increasing returns to scale*, and if the inequality is reversed the production function exhibits *decreasing returns to scale*.

[3] You may recall from microeconomics that with constant returns to scale the number of firms is indeterminate—i.e., not pinned down by the model.

FIGURE 2.1 **A COBB-DOUGLAS PRODUCTION FUNCTION**



The term on the left-hand side of equation (2.3) is the continuous time version of $K_{t+1} - K_t$, that is, the change in the capital stock per "period." We use the "dot" notation[4] to denote a derivative with respect to time:

$$\dot{K} \equiv \frac{dK}{dt}.$$

The second term of equation (2.3) represents gross investment. Following Solow, we assume that workers/consumers save a constant fraction, s, of their combined wage and rental income, $Y = wL + rK$. The economy is closed, so that saving equals investment, and the only use of investment in this economy is to accumulate capital. The consumers then rent this capital to firms for use in production, as discussed above.

The third term of equation (2.3) reflects the depreciation of the capital stock that occurs during production. The standard functional form used here implies that a constant fraction, d, of the capital stock depreciates every period (regardless of how much output is produced). For

[4]Appendix A discusses the meaning of this notation in more detail.

example, we often assume $d = .05$, so that 5 percent of the machines and factories in our model economy wear out each year.

To study the evolution of output per person in this economy, we rewrite the capital accumulation equation in terms of capital per person. Then the production function in equation (2.2) will tell us the amount of output per person produced for whatever capital stock per person is present in the economy. This rewriting is most easily accomplished by using a simple mathematical trick that is often used in the study of growth. The mathematical trick is to "take logs and then derivatives" (see Appendix A for further discussion). Two examples of this trick are given below.

Example 1:

$$k \equiv K/L \Longrightarrow \log k = \log K - \log L$$
$$\Longrightarrow \frac{\dot{k}}{k} = \frac{\dot{K}}{K} - \frac{\dot{L}}{L}.$$

Example 2:

$$y = k^{\alpha} \Longrightarrow \log y = \alpha \log k$$
$$\Longrightarrow \frac{\dot{y}}{y} = \alpha \frac{\dot{k}}{k}.$$

Applying Example 1 to equation (2.3) will allow us to rewrite the capital accumulation equation in terms of capital per worker. But before we proceed, let's first consider the growth rate of the labor force, $\dot{L}/L$. An important assumption that will be maintained throughout most of this book is that the labor force participation rate is constant and that the population growth rate is given by the parameter n.[5] This implies that the labor force growth rate, $\dot{L}/L$, is also given by n. If $n = .01$, then the population and the labor force are growing at one percent per year. This exponential growth can be seen from the relationship

$$L(t) = L_0 e^{nt}.$$

Take logs and differentiate this equation, and what do you get?

[5]Often, it is convenient in describing the model to assume that the labor force participation rate is unity—i.e., every member of the population is also a worker.

Now we are ready to combine Example 1 and equation (2.3):

$$\frac{\dot{k}}{k} = \frac{sY}{K} - n - d$$
$$= \frac{sy}{k} - n - d.$$

This now yields the capital accumulation equation in per worker terms:

$$\dot{k} = sy - (n + d)k.$$

This equation says that the change in capital per worker each period is determined by three terms. Two of the terms are analogous to the original capital accumulation equation. Investment per worker, sy, increases k, while depreciation per worker, dk, reduces k. The term that is new in this equation is a reduction in k because of population growth, the nk term. Each period, there are nL new workers around who were not there during the last period. If there were no new investment and no depreciation, capital *per worker* would decline because of the increase in the labor force. The amount by which it would decline is exactly nk, as can be seen by setting $\dot{K}$ to zero in Example 1.

2.1.1 SOLVING THE BASIC SOLOW MODEL

We have now laid out the basic elements of the Solow model and it is time to begin solving the model. What does it mean to "solve" a model? To answer this question we need to explain exactly what a model is and to define some concepts.

In general, a model consists of several equations that describe the relationships among a collection of *endogenous variables*—that is, among variables whose values are determined within the model itself. For example, equation (2.1) shows how output is produced from capital and labor, and equation (2.3) shows how capital is accumulated over time. Output, Y, and capital, K, are endogenous variables, as are the respective "per worker" versions of these variables, y and k.

Notice that the equations describing the relationships among endogenous variables also involve *parameters* and *exogenous variables*. Parameters are terms such as α, s, k_0, and n that stand in for single numbers. Exogenous variables are terms that may vary over time but whose values are determined outside of the model—i.e., exogenously.

The number of workers in this economy, L, is an example of an exogenous variable.

With these concepts explained, we are ready to tackle the question of what it means to solve a model. Solving a model means obtaining the values of each endogenous variable when given values for the exogenous variables and parameters. Ideally, one would like to be able to express each endogenous variable as a function only of exogenous variables and parameters. Sometimes this is possible; other times a diagram can provide insights into the nature of the solution but a computer is needed for exact values.

For this purpose, it is helpful to think of the economist as a laboratory scientist. The economist sets up a model and has control over the parameters and exogenous variables. The "experiment" is the model itself. Once the model is setup, the economist starts the experiment and watches to see how the endogenous variables evolve over time. The economist is free to vary the parameters and exogenous variables in different experiments to see how this changes the evolution of the endogenous variables.

In the case of the Solow model, our solution will proceed in several steps. We begin with several diagrams that provide insight into the solution. Then, in Section 2.1.4, we provide an analytic solution for the long-run values of the key endogenous variables. A full solution of the model at every point in time is possible analytically, but this derivation is somewhat difficult and is relegated to the appendix of this chapter.

2.1.2 THE SOLOW DIAGRAM

At the beginning of this section we derived the two key equations of the Solow model in terms of output per worker and capital per worker. These equations are

$$y = k^{\alpha} \tag{2.4}$$

and

$$\dot{k} = sy - (n + d)k. \tag{2.5}$$

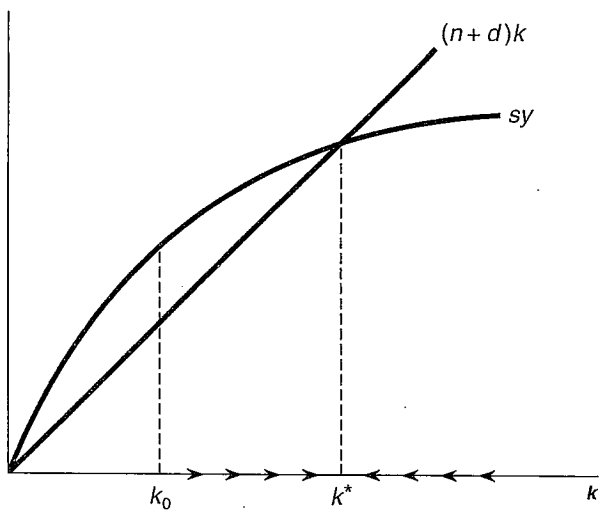
Now we are ready to ask fundamental questions of our model. For example, an economy starts out with a given stock of capital per worker, k_0, and a given population growth rate, depreciation rate, and investment

rate. How does output per worker evolve over time in this economy — i.e., how does the economy grow? How does output per worker compare in the long run between two economies that have different investment rates?

These questions are most easily analyzed in a Solow diagram, as shown in Figure 2.2. The Solow diagram consists of two curves, plotted as functions of the capital-labor ratio, k. The first curve is the amount of investment per person, $sy = sk^{\alpha}$. This curve has the same shape as the production function plotted in Figure 2.1, but it is translated down by the factor s. The second curve is the line $(n + d)k$, which represents the amount of new investment per person required to keep the amount of capital per worker constant — both depreciation and the growing workforce tend to reduce the amount of capital per person in the economy. By no coincidence, the difference between these two curves is the change in the amount of capital per worker. When this change is positive and the economy is increasing its capital per worker, we say that *capital deepening* is occurring. When this per worker change is zero but the actual capital stock K is growing (because of population growth), we say that only *capital widening* is occurring.

To consider a specific example, suppose an economy has capital equal to the amount k_0 today, as drawn in Figure 2.2. What happens over

FIGURE 2.2 **THE BASIC SOLOW DIAGRAM**



time? At k_0, the amount of investment per worker exceeds the amount needed to keep capital per worker constant, so that capital deepening occurs — that is, k increases over time. This capital deepening will continue until $k = k^*$, at which point $sy = (n + d)k$, so that $\dot{k} = 0$. At this point, the amount of capital per worker remains constant, and we call such a point a *steady state*.

What would happen if instead the economy began with a capital stock per worker larger than k^*? At points to the right of k^* in Figure 2.2, the amount of investment per worker provided by the economy is less than the amount needed to keep the capital-labor ratio constant. The term $\dot{k}$ is negative, and therefore the amount of capital per worker begins to decline in this economy. This decline occurs until the amount of capital per worker falls to k^*.

Notice that the Solow diagram determines the steady-state value of capital per worker. The production function of equation (2.4) then determines the steady-state value of output per worker, y^*, as a function of k^*. It is sometimes convenient to include the production function in the Solow diagram itself to make this point clearly. This is done in

FIGURE 2.3 **THE SOLOW DIAGRAM AND THE PRODUCTION FUNCTION**

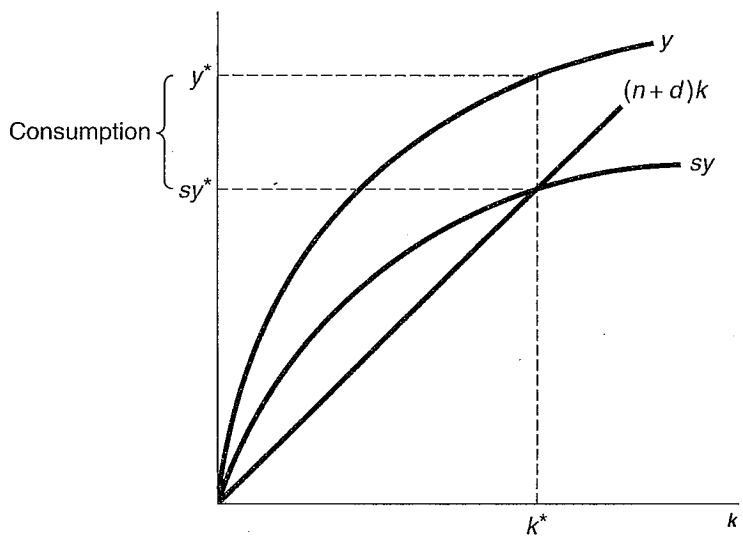


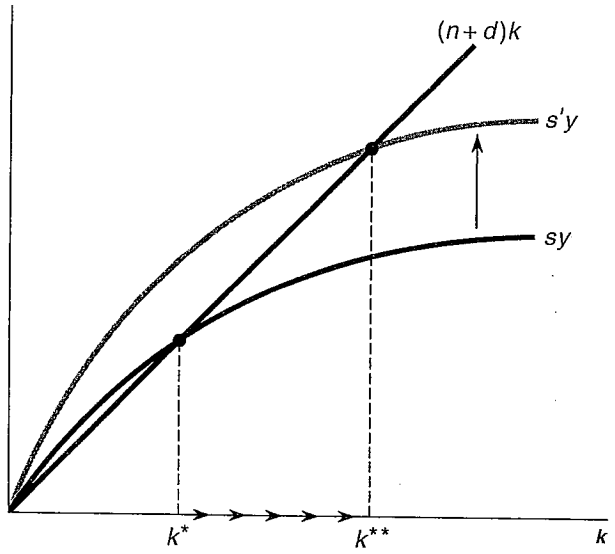
Figure 2.3. Notice that steady-state consumption per worker is then given by the difference between steady-state output per worker, y^*, and steady-state investment per worker, sy^*.

2.1.3 COMPARATIVE STATICS

Comparative statics are used to examine the response of the model to changes in the values of various parameters. In this section, we will consider what happens to per capita income in an economy that begins in steady state but then experiences a "shock." The shocks we will consider are an increase in the investment rate, s, and an increase in the population growth rate, n.

AN INCREASE IN THE INVESTMENT RATE Consider an economy that has arrived at its steady-state value of output per worker. Now suppose that the consumers in that economy decide to increase the investment rate permanently from s to some value s'. What happens to k and y in this economy?

FIGURE 2.4 AN INCREASE IN THE INVESTMENT RATE

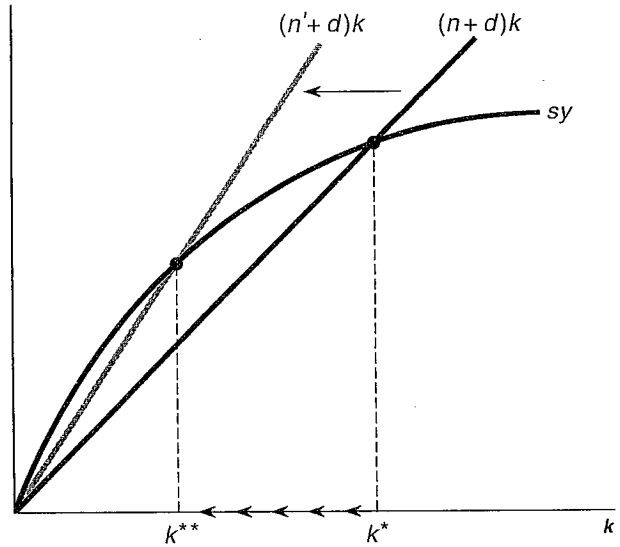


The answer is found in Figure 2.4. The increase in the investment rate shifts the sy curve upward to $s'y$. At the current value of the capital stock, k^*, investment per worker now exceeds the amount required to keep capital per worker constant, and therefore the economy begins capital deepening again. This capital deepening continues until $s'y = (n + d)k$ and the capital stock per worker reaches a higher value, indicated by the point k^{**}. From the production function, we know that this higher level of capital per worker will be associated with higher per capita output; the economy is now richer than it was before.

AN INCREASE IN THE POPULATION GROWTH RATE Now consider an alternative exercise. Suppose an economy has reached its steady state, but then because of immigration, for example, the population growth rate of the economy rises from n to n'. What happens to k and y in this economy?

Figure 2.5 computes the answer graphically. The $(n + d)k$ curve rotates up and to the left to the new curve $(n' + d)k$. At the current value of the capital stock, k^*, investment per worker is now no longer high enough to keep the capital-labor ratio constant in the face of the rising

FIGURE 2.5 AN INCREASE IN POPULATION GROWTH



population. Therefore the capital-labor ratio begins to fall. It continues to fall until the point at which $sy = (n' + d)k$, indicated by k^{**} in Figure 2.5. At this point, the economy has less capital per worker than it began with and is therefore poorer: per capita output is ultimately lower after the increase in population growth in this example. Why?

2.1.4 PROPERTIES OF THE STEADY STATE

By definition, the steady-state quantity of capital per worker is determined by the condition that $\dot{k} = 0$. Equations (2.4) and (2.5) allow us to use this condition to solve for the steady-state quantities of capital per worker and output per worker. Substituting from (2.4) into (2.5),

$$\dot{k} = sk^{\alpha} - (n + d)k,$$

and setting this equation equal to zero yields

$$k^* = \left(\frac{s}{n + d}\right)^{1/(1-\alpha)}.$$

Substituting this into the production function reveals the steady-state quantity of output per worker, y^*:

$$y^* = \left(\frac{s}{n + d}\right)^{\alpha/(1-\alpha)}.$$

Notice that the endogenous variable y^* is now written in terms of the parameters of the model. Thus, we have a "solution" for the model, at least in the steady state.

This equation reveals the Solow model's answer to the question "Why are we so rich and they so poor?" Countries that have high savings/investment rates will tend to be richer, *ceteris paribus*.[6] Such countries accumulate more capital per worker, and countries with more capital per worker have more output per worker. Countries that have high population growth rates, in contrast, will tend to be poorer, according to the Solow model. A higher fraction of savings in these economies must go simply to keep the capital-labor ratio constant in the face of a growing population. This capital-widening requirement makes capital

[6] *Ceteris paribus* is Latin for "all other things being equal."

deepening more difficult, and these economies tend to accumulate less capital per worker.

How well do these predictions of the Solow model hold up empirically? Figures 2.6 and 2.7 plot GDP per worker against gross investment as a share of GDP and against population growth rates, respectively. Broadly speaking, the predictions of the Solow model are borne out by the empirical evidence. Countries with high investment rates tend to be richer on average than countries with low investment rates, and countries with high population growth rates tend to be poorer on average. At this level, then, the general predictions of the Solow model seem to be supported by the data.

FIGURE 2.6 GDP PER WORKER VERSUS THE INVESTMENT RATE

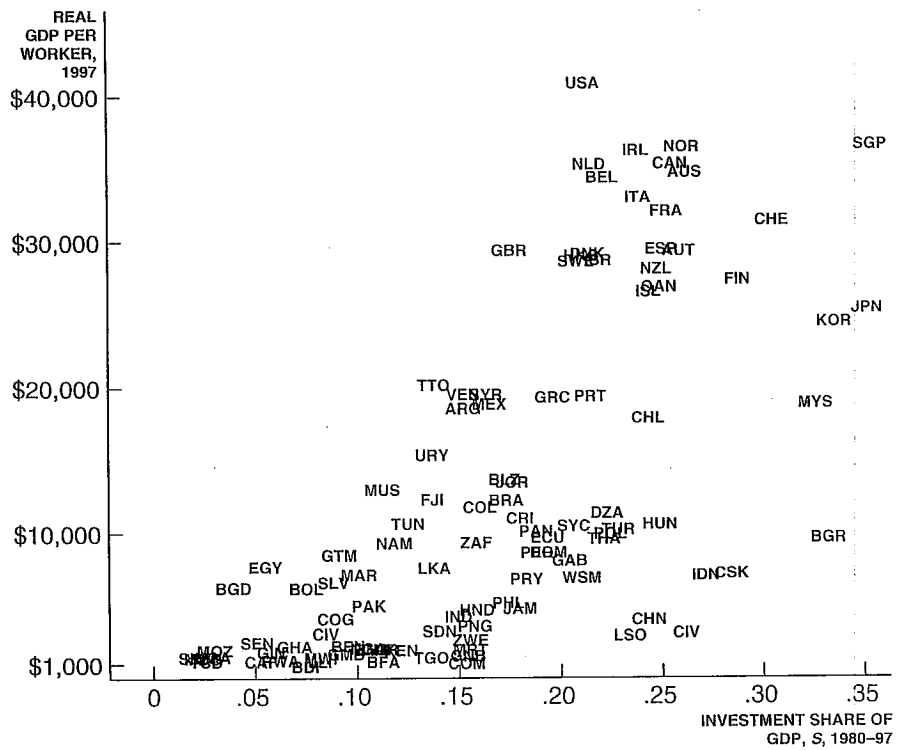
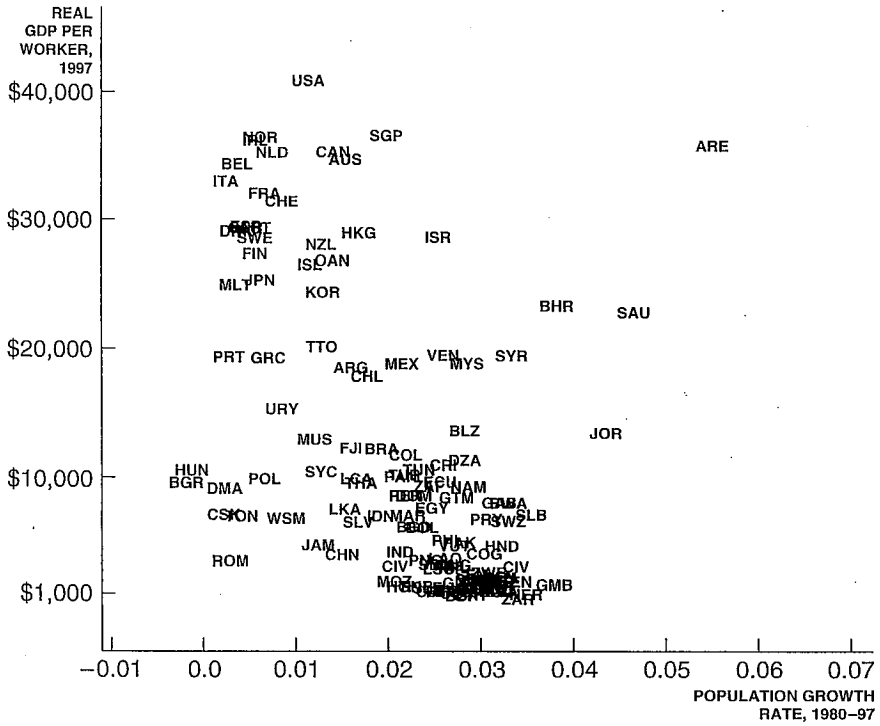


FIGURE 2.7 GDP PER WORKER VERSUS POPULATION GROWTH RATES



2.1.5 ECONOMIC GROWTH IN THE SIMPLE MODEL

What does economic growth look like in the steady state of this simple version of the Solow model? The answer is that there is *no* per capita growth in this version of the model! Output per worker (and therefore per person, since we've assumed the labor force participation rate is constant) is constant in the steady state. Output itself, Y, is growing, of course, but only at the rate of population growth.[7]

This version of the model fits several of the stylized facts discussed in Chapter 1. It generates differences in per capita income across countries. It generates a constant capital-output ratio (because both k and y are constant, implying that K/Y is constant). It generates a constant interest rate, the marginal product of capital. However, it fails to predict a very important stylized fact: that economies exhibit sustained per capita income growth. In this model, economies may grow for a while, but not forever. For example, an economy that begins with a stock of capital per worker below its steady-state value will experience growth in k and y along the *transition path* to the steady state. Over time, however, growth slows down as the economy approaches its steady state, and eventually growth stops altogether.

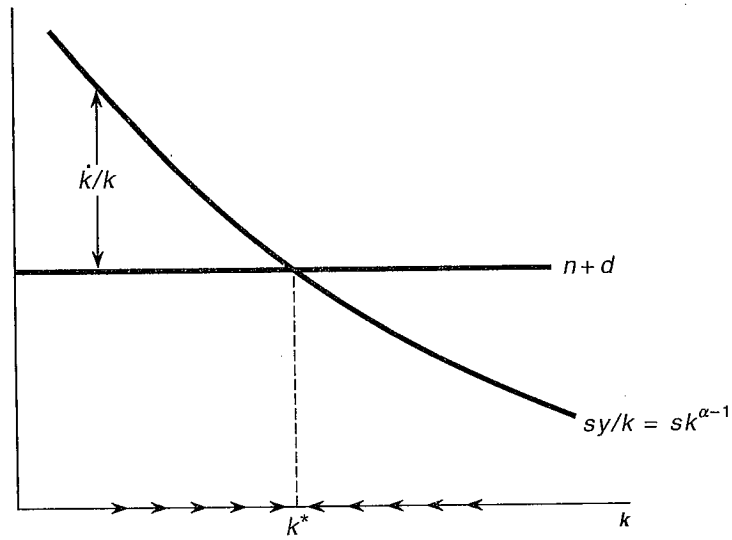
[7]This can be seen easily by applying the "take logs and differentiate" trick to $y \equiv Y/L$.

To see that growth slows down along the transition path, notice two things. First, from the capital accumulation equation (equation (2.5)), one can divide both sides by k to get

$$\frac{\dot{k}}{k} = sk^{\alpha-1} - (n + d). \tag{2.6}$$

Because α is less than one, as k rises, the growth rate of k gradually declines. Second, from Example 2, the growth rate of y is proportional to the growth rate of k, so that the same statement holds true for output per worker.

FIGURE 2.8 TRANSITION DYNAMICS



The transition dynamics implied by equation (2.6) are plotted in Figure 2.8.[8] The first term on the right-hand side of the equation is $sk^{\alpha-1}$, which is equal to sy/k. The higher the level of capital per worker, the lower the average product of capital, y/k, because of diminishing returns to capital accumulation (α is less than one). Therefore, this curve slopes downward. The second term on the right-hand side of equation (2.6) is $n + d$, which doesn't depend on k, so it is plotted as a horizontal line. The difference between the two lines in Figure 2.8 is the growth rate of the capital stock, or $\dot{k}/k$. Thus, the figure clearly indicates that the further an economy is below its steady-state value of k, the faster the economy grows. Also, the further an economy is above its steady-state value of k, the faster k declines.

2.2 TECHNOLOGY AND THE SOLOW MODEL

To generate sustained growth in per capita income in this model, we must follow Solow and introduce technological progress to the model. This is accomplished by adding a technology variable, A, to the production function:

$$Y = F(K, AL) = K^{\alpha}(AL)^{1-\alpha}. \tag{2.7}$$

Entered this way, the technology variable A is said to be "labor-augmenting" or "Harrod-neutral."[9] Technological progress occurs when A increases over time — a unit of labor, for example, is more productive when the level of technology is higher.

An important assumption of the Solow model is that technological progress is *exogenous*: in a common phrase, technology is like "manna from heaven," in that it descends upon the economy automatically and regardless of whatever else is going on in the economy. Instead of modeling carefully where technology comes from, we simply recognize for the moment that there is technological progress and make the assumption that A is growing at a constant rate:

$$\frac{\dot{A}}{A} = g \Longleftrightarrow A = A_0 e^{gt},$$

where g is a parameter representing the growth rate of technology. Of course, this assumption about technology is unrealistic, and explaining how to relax this assumption is one of the major accomplishments of the "new" growth theory that we will explore in later chapters.

The capital accumulation equation in the Solow model with technology is the same as before. Rewriting it slightly, we get

$$\frac{\dot{K}}{K} = s\frac{Y}{K} - d. \tag{2.8}$$

To see the growth implications of the model with technology, first rewrite the production function (2.7) in terms of output per worker:

$$y = k^{\alpha}A^{1-\alpha}.$$

Then take logs and differentiate:

$$\frac{\dot{y}}{y} = \alpha\frac{\dot{k}}{k} + (1-\alpha)\frac{\dot{A}}{A}. \tag{2.9}$$

Finally, notice from the capital accumulation equation (2.8) that the growth rate of K will be constant if and only if Y/K is constant. Furthermore, if Y/K is constant, y/k is also constant, and most important, y and k will be growing at the same rate. A situation in which capital, output, consumption, and population are growing at constant rates is called a *balanced growth path*. Partly because of its empirical appeal, this is a situation that we often wish to analyze in our models. For example, according to Fact 5 in Chapter 1, this situation describes the U.S. economy.

Let's use the notation g_x to denote the growth rate of some variable x along a balanced growth path. Then, along a balanced growth path, $g_y = g_k$ according to the argument above. Substituting this relationship into equation (2.9) and recalling that $\dot{A}/A = g$,

$$g_y = g_k = g. \tag{2.10}$$

That is, along a balanced growth path in the Solow model, output per worker and capital per worker both grow at the rate of exogenous tech-

[8] This alternative version of the Solow diagram makes the growth implications of the Solow model much more transparent. Xavier Sala-i-Martin (1990) emphasizes this point.

[9] The other possibilities are $F(AK, L)$, which is known as "capital-augmenting" or "Solow-neutral" technology, and $AF(K, L)$, which is known as "Hicks-neutral" technology. With the Cobb-Douglas functional form assumed here, this distinction is less important.

nological change, g. Notice that in the model of Section 2.1, there was no technological progress, and therefore there was no long-run growth in output per worker or capital per worker; $g_y = g_k = g = 0$. The model with technology reveals that *technological progress is the source of sustained per capita growth.* In this chapter, this result is little more than an assumption; in later chapters, we will explore the result in much more detail and come to the same conclusion.

2.2.1 THE SOLOW DIAGRAM WITH TECHNOLOGY

The analysis of the Solow model with technological progress proceeds very much like the analysis in Section 2.1: we set up a differential equation and analyze it in a Solow diagram to find the steady state. The only important difference is that the variable k is no longer constant in the long run, so we have to write our differential equation in terms of another variable. The new *state* variable will be $\tilde{k} \equiv K/AL$. Notice that this is equivalent to k/A and is obviously constant along the balanced growth path because $g_k = g_A = g$. The variable $\tilde{k}$ therefore represents the ratio of capital per worker to technology. We will refer to this as the "capital-technology" ratio (keeping in mind that the numerator is capital per worker rather than the total level of capital).

Rewriting the production function in terms of $\tilde{k}$, we get

$$\tilde{y} = \tilde{k}^{\alpha}, \tag{2.11}$$

where $\tilde{y} \equiv Y/AL = y/A$. Following the terminology above, we will refer to $\tilde{y}$ as the "output-technology ratio."[10]

Rewriting the capital accumulation equation in terms of $\tilde{k}$ is accomplished by following exactly the methodology used in Section 2.1. First, note that

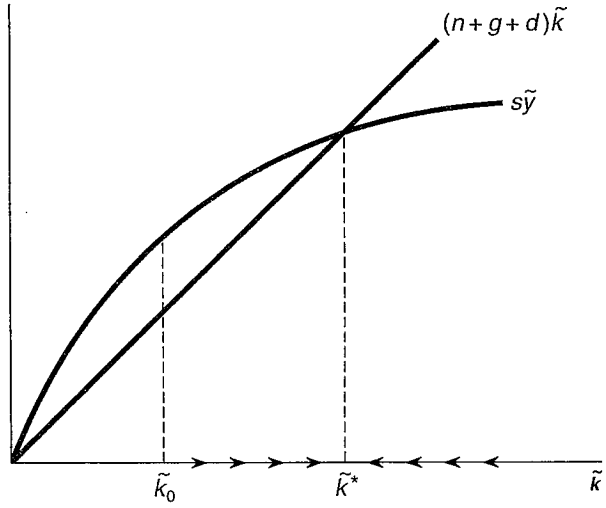
$$\frac{\dot{\tilde{k}}}{\tilde{k}} = \frac{\dot{K}}{K} - \frac{\dot{A}}{A} - \frac{\dot{L}}{L}.$$

Combining this with the capital accumulation equation reveals that

$$\dot{\tilde{k}} = s\tilde{y} - (n + g + d)\tilde{k}. \tag{2.12}$$

[10]The variables $\tilde{y}$ and $\tilde{k}$ are sometimes referred to as "output per effective unit of labor" and "capital per effective unit of labor." This labeling is motivated by the fact that technological progress is labor-augmenting. AL is then the "effective" amount of labor used in production.

FIGURE 2.9 THE SOLOW DIAGRAM WITH TECHNOLOGICAL PROGRESS



The similarity of equations (2.11) and (2.12) to their counterparts in Section 2.1 should be obvious.

The Solow diagram with technological progress is presented in Figure 2.9. The analysis of this diagram is very similar to the analysis when there is no technological progress, but the interpretation is slightly different. If the economy begins with a capital-technology ratio that is below its steady-state level, say at a point such as $\tilde{k}_0$, the capital-technology ratio will rise gradually over time. Why? Because the amount of investment being undertaken exceeds the amount needed to keep the capital-technology ratio constant. This will be true until $s\tilde{y} = (n+g+d)\tilde{k}$ at the point $\tilde{k}^*$, at which point the economy is in steady state and grows along a balanced growth path.

2.2.2 SOLVING FOR THE STEADY STATE

The steady-state output-technology ratio is determined by the production function and the condition that $\dot{\tilde{k}} = 0$. Solving for $\tilde{k}^*$, we find

that

$$\tilde{k}^* = \left(\frac{s}{n+g+d}\right)^{1/(1-\alpha)}.$$

Substituting this into the production function yields

$$\tilde{y}^* = \left(\frac{s}{n+g+d}\right)^{\alpha/(1-\alpha)}.$$

To see what this implies about output per worker, rewrite the equation as

$$y^*(t) = A(t)\left(\frac{s}{n+g+d}\right)^{\alpha/(1-\alpha)}, \tag{2.13}$$

where we explicitly note the dependence of y and A on time. From equation (2.13), we see that output per worker along the balanced growth path is determined by technology, the investment rate, and the population growth rate. For the special case of $g = 0$ and $A_0 = 1$—i.e., of no technological progress—this result is identical to that derived in Section 2.1.

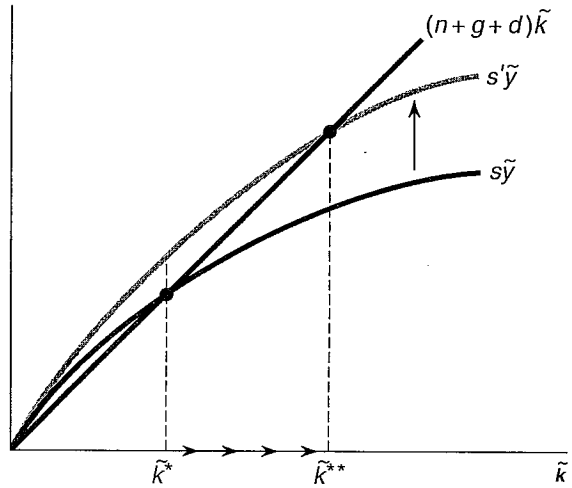
An interesting result is apparent from equation (2.13) and is discussed in more detail in Exercise 1 at the end of this chapter. That is, changes in the investment rate or the population growth rate affect the long-run *level* of output per worker but do not affect the long-run *growth rate* of output per worker. To see this more clearly, let's consider a simple example.

Suppose an economy begins in steady state with investment rate s and then permanently increases its investment rate to s' (for example, because of a permanent subsidy to investment). The Solow diagram for this policy change is drawn in Figure 2.10, and the results are broadly similar to the case with no technological progress. At the initial capital-technology ratio $\tilde{k}^*$, investment exceeds the amount needed to keep the capital-technology ratio constant, so $\tilde{k}$ begins to rise.

To see the effects on growth, rewrite equation (2.12) as

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = s\frac{\tilde{y}}{\tilde{k}} - (n+g+d),$$

FIGURE 2.10 AN INCREASE IN THE INVESTMENT RATE



and note that $\tilde{y}/\tilde{k}$ is equal to $\tilde{k}^{\alpha-1}$. Figure 2.11 illustrates the transition dynamics implied by this equation. As the diagram shows, the increase in the investment rate to s' raises the growth rate temporarily as the economy transits to the new steady state, $\tilde{k}^{**}$. Since g is constant, faster growth in $\tilde{k}$ along the transition path implies that output per worker increases more rapidly than technology: $\dot{y}/y > g$. The behavior of the growth rate of output per worker over time is displayed in Figure 2.12.

Figure 2.13 cumulates the effects on growth to show what happens to the (log) level of output per worker over time. Prior to the policy change, output per worker is growing at the constant rate g, so that the log of output per worker rises linearly. At the time of the policy change, t^*, output per worker begins to grow more rapidly. This more rapid growth continues temporarily until the output-technology ratio reaches its new steady state. At this point, growth has returned to its long-run level of g.

This exercise illustrates two important points. First, policy changes in the Solow model increase growth rates, but only temporarily along the transition to the new steady state. That is, policy changes have no long-run *growth effect*. Second, policy changes can have *level effects*.

FIGURE 2.11 AN INCREASE IN THE INVESTMENT RATE: TRANSITION DYNAMICS

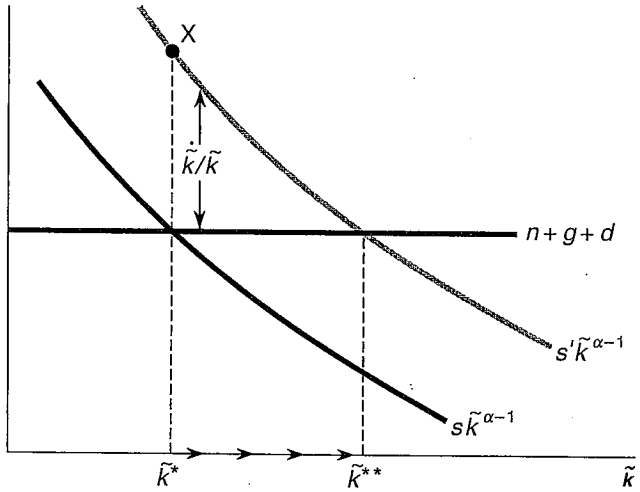


FIGURE 2.12 THE EFFECT OF AN INCREASE IN INVESTMENT ON GROWTH

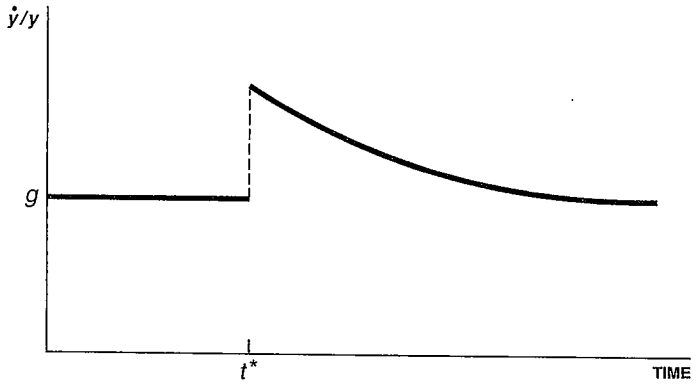
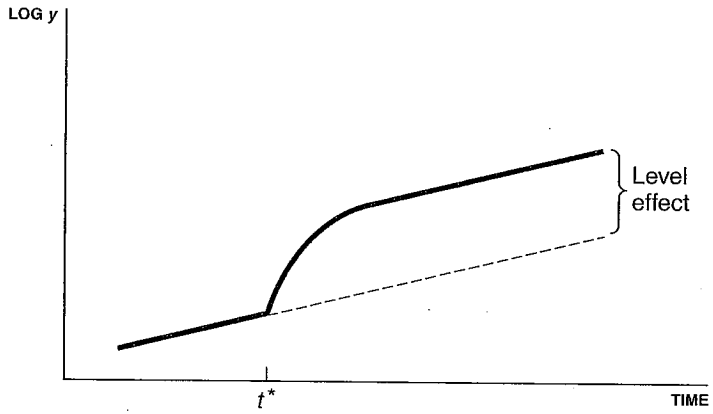


FIGURE 2.13 THE EFFECT OF AN INCREASE IN INVESTMENT ON y



That is, a permanent policy change can permanently raise (or lower) the level of per capita output.

2.3 EVALUATING THE SOLOW MODEL

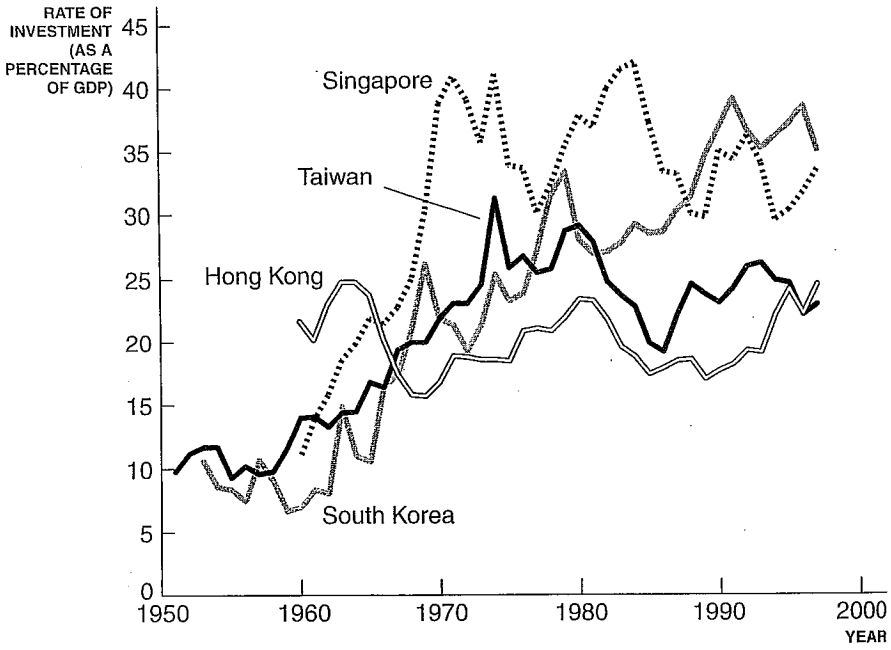
How does the Solow model answer the key questions of growth and development? First, the Solow model appeals to differences in investment rates and population growth rates and (perhaps) to exogenous differences in technology to explain differences in per capita incomes. Why are we so rich and they so poor? According to the Solow model, it is because we invest more and have lower population growth rates, both of which allow us to accumulate more capital per worker and thus increase labor productivity. In the next chapter, we will explore this hypothesis more carefully and see that it is firmly supported by data across the countries of the world.

Second, why do economies exhibit sustained growth in the Solow model? The answer is technological progress. As we saw earlier, without technological progress, per capita growth will eventually cease as diminishing returns to capital set in. Technological progress, however, can offset the tendency for the marginal product of capital to fall, and

in the long run, countries exhibit per capita growth at the rate of technological progress.

How, then, does the Solow model account for differences in growth rates across countries? At first glance, it may seem that the Solow model cannot do so, except by appealing to differences in (unmodeled) technological progress. A more subtle explanation, however, can be found by appealing to transition dynamics. We have seen several examples of how transition dynamics can allow countries to grow at rates different from their long-run growth rates. For example, an economy with a capital-technology ratio below its long-run level will grow rapidly until the capital-technology ratio reaches its steady-state level. This reasoning may help explain why countries such as Japan and Germany, which had their capital stocks wiped out by World War II, have grown more rapidly than the United States over the last fifty years. Or it may explain why an economy that increases its investment rate will grow rapidly as

FIGURE 2.14 **INVESTMENT RATES IN SOME NEWLY INDUSTRIALIZING ECONOMIES**



it makes the transition to a higher output-technology ratio. This explanation may work well for countries such as South Korea, Singapore, and Taiwan. Their investment rates have increased dramatically since 1950, as shown in Figure 2.14. The explanation may work less well, however, for an economy such as Hong Kong's. This kind of reasoning raises an interesting question: can countries permanently grow at different rates? This question will be discussed in more detail in later chapters.

2.4 GROWTH ACCOUNTING, THE PRODUCTIVITY SLOWDOWN, AND THE NEW ECONOMY

We have seen in the Solow model that sustained growth occurs only in the presence of technological progress. Without technological progress, capital accumulation runs into diminishing returns. With technological progress, however, improvements in technology continually offset the diminishing returns to capital accumulation. Labor productivity grows as a result, both directly because of the improvements in technology and indirectly because of the additional capital accumulation these improvements make possible.

In 1957, Solow published a second article, "Technical Change and the Aggregate Production Function," in which he performed a simple accounting exercise to break down growth in output into growth in capital, growth in labor, and growth in technological change. This "growth-accounting" exercise begins by postulating a production function such as

$$Y = BK^{\alpha}L^{1-\alpha},$$

where B is a Hicks-neutral productivity term.[11] Taking logs and differentiating this production function, one derives the key formula of growth accounting:

$$\frac{\dot{Y}}{Y} = \alpha\frac{\dot{K}}{K} + (1-\alpha)\frac{\dot{L}}{L} + \frac{\dot{B}}{B}. \tag{2.14}$$

[11] In fact, this growth accounting can be done with a much more general production function such as $B(t)F(K,L)$, and the results are very similar.

This equation says that output growth is equal to a weighted average of capital and labor growth plus the growth rate of B. This last term, $\dot{B}/B$, is commonly referred to as *total factor productivity growth* or *multifactor productivity growth*. Solow, as well as economists such as Edward Denison and Dale Jorgenson who followed Solow's approach, have used this equation to understand the sources of growth in output.

Since we are primarily interested here in the growth rate of output per worker instead of total output, it is helpful to rewrite equation (2.14) by subtracting $\dot{L}/L$ from both sides:

$$\frac{\dot{y}}{y} = \alpha\frac{\dot{k}}{k} + \frac{\dot{B}}{B}. \tag{2.15}$$

That is, the growth rate of output per worker is decomposed into the contribution of physical capital per worker and the contribution from multifactor productivity growth.

The U.S. Bureau of Labor Statistics (BLS) provides a detailed accounting of U.S. growth using a generalization of equation (2.15). Its most recent numbers are reported in Table 2.1. They generalize this equation in a couple of ways. First, the BLS measures labor by calculating total hours worked rather than just the number of workers. Second, the BLS includes an additional term in equation (2.15) to adjust for the changing composition of the labor force—to recognize, for example, that the labor force is more educated today than it was forty years ago.

TABLE 2.1 GROWTH ACCOUNTING FOR THE UNITED STATES

	1948–98	48–73	73–79	79–90	90–95	95–98
Output per hour	2.5	3.3	1.3	1.6	1.5	2.5
Contributions from:						
Capital per hour worked	0.8	1.0	0.7	0.7	0.5	0.8
Information technology	0.3	0.1	0.3	0.5	0.4	0.8
Other capital services	0.6	0.9	0.5	0.3	0.1	0.0
Labor composition	0.2	0.2	0.0	0.3	0.4	0.3
Multifactor productivity	1.4	2.1	0.6	0.5	0.6	1.4

SOURCE: Bureau of Labor Statistics (2000).
Note: The table reports average annual growth rates for the private business sector. "Information technology" refers to information processing equipment and software.

As can be seen from the table, output per hour in the private business sector for the United States grew at an average annual rate of 2.5 percent between 1948 and 1998. The contribution from capital per hour worked was 0.8 percentage points, and the changing composition of the labor force contributed another 0.2 percentage points. Multifactor productivity growth accounts for the remaining 1.4 percentage points, by definition. The implication is that about one-half of U.S. growth was due to factor accumulation and one-half was due to the improvement in the productivity of these factors over this period. Because of the way in which it is calculated, economists have referred to this 1.4 percent as the "residual" or even as a "measure of our ignorance." One interpretation of the multifactor productivity growth term is that it is due to technological change; notice that in terms of the production function in equation (2.7), $B = A^{1-\alpha}$. This interpretation will be explored in later chapters.

Table 2.1 also reveals how GDP growth and its sources have changed over time in the United States. One of the important stylized facts revealed in the table is the productivity growth slowdown that occurred in the 1970s. The top row shows that growth in output per hour (also known as labor productivity) slowed dramatically after 1973; growth between 1973 and 1995 was nearly 2 percentage points slower than growth between 1948 and 1973. What was the source of this slowdown? The next few rows show that the changes in the contributions from capital per worker and labor composition are relatively minor. The primary culprit of the productivity slowdown is a substantial decline in the growth rate of multifactor productivity. For some reason, the "residual" was much lower after 1973 than before: the bulk of the productivity slowdown is accounted for by the "measure of our ignorance." A similar productivity slowdown occurred throughout the advanced countries of the world.

Various explanations for the productivity slowdown have been advanced. For example, perhaps the sharp rise in energy prices in 1973 and 1979 contributed to the slowdown. One problem with this explanation is that in real terms energy prices were lower in the late 1980s

than they were before the oil shocks. Another explanation may involve the changing composition of the labor force or the sectoral shift in the economy away from manufacturing (which tends to have high labor productivity) toward services (many of which have low labor productivity). This explanation receives some support from recent evidence that productivity growth recovered substantially in the 1980s in manufacturing. It is possible that a slowdown in resources spent on research and development (R&D) in the late 1960s contributed to the slowdown as well. Or, perhaps it is not the 1970s and 1980s that need to be explained but rather the 1950s and 1960s: growth may simply have been artificially and temporarily high in the years following World War II because of the application to the private sector of new technologies created for the war. Nevertheless, careful work on the productivity slowdown has failed to provide a complete explanation.[12]

The flip side of the productivity slowdown after 1973 is the rise in productivity growth in the 1995–98 period, sometimes labeled the "New Economy." Growth in output per hour and in multifactor productivity rose substantially in this period, returning about 50 percent of the way back to the growth rates exhibited before 1973. As shown in Table 2.1, the increase in growth rates is partially associated with an increase in the use of information technology. Before 1973, this component of capital accumulation contributed only 0.1 percentage points of growth, but by the late 1990s, this contribution had risen to 0.8 percentage points. In addition, evidence suggests that as much as half of the rise in multifactor productivity growth in recent years is due to increases in efficiency of the production of information technology.

Recently, a number of economists have suggested that the information-technology revolution associated with the widespread adoption of computers might explain both the productivity slowdown after 1973 as well as the recent rise in productivity growth. According to this hypothesis, growth slowed temporarily while the economy adapted its factories to the new production techniques associated with information technology and as workers learned to take advantage of the new technology. The recent upsurge in productivity growth, then, reflects the

[12]The fall 1988 issue of the *Journal of Economic Perspectives* contains several papers discussing potential explanations of the productivity slowdown.

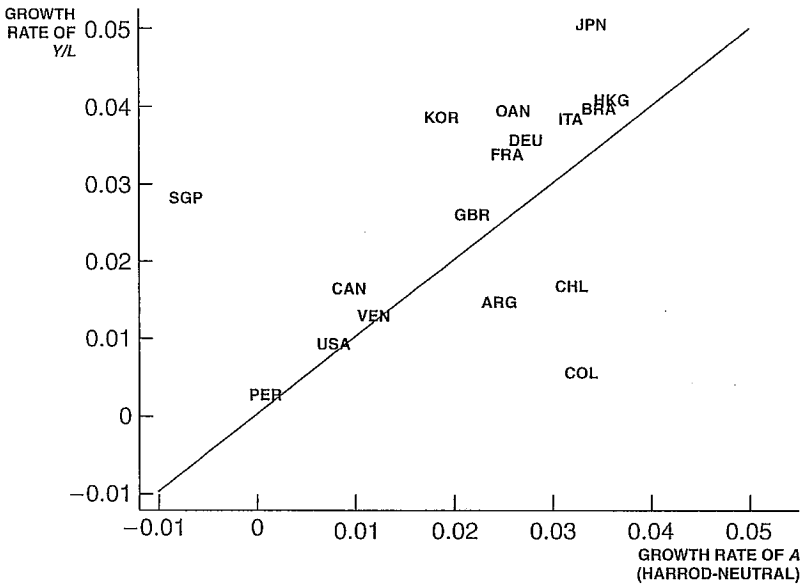
successful widespread adoption of this new technology. The recent upsurge in productivity growth, then, reflects the successful widespread adoption of this new technology.[13] Whether or not this view is correct remains to be seen.

Growth accounting has also been used to analyze economic growth in countries other than the United States. One of the more interesting applications is to the NICs of South Korea, Hong Kong, Singapore, and Taiwan. Recall from Chapter 1 that average annual growth rates have exceeded 5 percent in these economies since 1960. Alwyn Young (1995) shows that a large part of this growth is the result of factor accumulation: increases in investment in physical capital and education, increases in labor force participation, and a shift from agriculture into manufacturing. Support for Young's result is provided in Figure 2.15. The vertical axis measures growth in output per worker, while the horizontal axis measures growth in Harrod-neutral (i.e., labor-augmenting) total factor productivity. That is, instead of focusing on growth in B, where $B = A^{1-\alpha}$, we focus on the growth of A. (Notice that with $\alpha = \frac{1}{3}$, the growth rate of A is simply 1.5 times the growth rate of B.) This change of variables is often convenient because along a steady-state balanced growth path, $g_y = g_A$. Countries growing along a balanced growth path, then, should lie on the 45-degree line in the figure.

Two features of Figure 2.15 stand out. First, while the growth rates of output per worker in the East Asian countries are clearly remarkable, their rates of growth in total factor productivity (TFP) are less so. A number of other countries such as Italy, Brazil, and Chile have also experienced rapid TFP growth. Total factor productivity growth, while typically higher than in the United States, was not exceptional in the East Asian economies. Second, the East Asian countries are far above the 45-degree line. This shift means that growth in output per worker is much higher than TFP growth would suggest. Singapore is an extreme example, with slightly *negative* TFP growth. Its rapid growth of output per worker is entirely attributable to growth in capital and education. More generally, a key source of the rapid growth performance

[13]See Paul David (1990) and Jeremy Greenwood and Mehmet Yorukoglu (1997). More generally, a nice collection of papers on the "New Economy" can be found in the Fall 2000 issue of the *Journal of Economic Perspectives*.

FIGURE 2.15 GROWTH ACCOUNTING



SOURCE: Author's calculations using the data collection reported in Table 10.8 of Barro and Sala-i-Martin (1998).

Note: The years over which growth rates are calculated vary across countries: 1960–90 for OECD members, 1940–80 for Latin America, and 1966–90 for East Asia.

of these countries is factor accumulation. Therefore, Young concludes, the framework of the Solow model (and the extension of the model in Chapter 3) can explain a substantial amount of the rapid growth of the East Asian economies.

APPENDIX: CLOSED-FORM SOLUTION OF THE SOLOW MODEL

It is possible to solve analytically for output per worker $y(t)$ at each point in time in the Solow model. The derivation of this solution is beyond the scope of this book. One derivation can be found in the appendix to

Chapter 1 of Barro and Sala-i-Martin (1998). Another can be found in "A Note on the Closed-Form Solution of the Solow Model," which can be downloaded from my Web page at http://emlab.berkeley.edu/users/chad/papers.html#closed form. The key insight is to recognize that the differential equation for the capital-output ratio in the Solow model is linear and can be solved using standard techniques.

Although the method of solution is beyond the scope of this book, the exact solution is still of interest. It illustrates nicely what it means to "solve" a model:

$$y(t) = \left(\frac{s}{n+g+d}(1 - e^{-\lambda t}) + \left(\frac{y_0}{A_0}\right)^{\frac{1-\alpha}{\alpha}} e^{-\lambda t} \right)^{\frac{\alpha}{1-\alpha}} A(t).$$

In this expression, we have defined a new parameter: $\lambda \equiv (1-\alpha)(n+g+d)$. Notice that output per worker at any time t is written as a function of the parameters of the model as well as of the exogenous variable $A(t)$.

To interpret this expression, notice that at $t = 0$, output per worker is simply equal to y_0, which in turn is given by the parameters of the model; recall that $y_0 = k_0^{\alpha} A_0^{1-\alpha}$. That's a good thing: our solution says that output per worker starts at the level given by the production function! At the other extreme, consider what happens as t gets very large, in the limit going off to infinity. In this case, $e^{-\lambda t}$ goes to zero, so we are left with an expression that is exactly that given by equation (2.13): output per worker reaches its steady-state value.

In between $t = 0$ and $t = \infty$, output per worker is some kind of weighted average of its initial value and its steady-state value. As time goes on, all that changes are the weights.

The interested reader will find it very useful to go back and reinterpret the Solow diagram and the various comparative static exercises with this solution in mind.

EXERCISES

1. *A decrease in the investment rate.* Suppose the U.S. Congress enacts legislation that discourages saving and investment, such as the elimination of the investment tax credit that occurred in 1990. As a result, suppose the investment rate falls permanently from s' to s''.

Examine this policy change in the Solow model with technological progress, assuming that the economy begins in steady state. Sketch a graph of how (the natural log of) output per worker evolves over time with and without the policy change. Make a similar graph for the growth rate of output per worker. Does the policy change permanently reduce the *level* or the *growth rate* of output per worker?

2. *An increase in the labor force.* Shocks to an economy, such as wars, famines, or the unification of two economies, often generate large one-time flows of workers across borders. What are the short-run and long-run effects on an economy of a one-time permanent increase in the stock of labor? Examine this question in the context of the Solow model with $g = 0$ and $n > 0$.

3. *An income tax.* Suppose the U.S. Congress decides to levy an income tax on both wage income and capital income. Instead of receiving $wL + rK = Y$, consumers receive $(1 - \tau)wL + (1 - \tau)rK = (1 - \tau)Y$. Trace the consequences of this tax for output per worker in the short and long runs, starting from steady state.

4. *Manna falls faster.* Suppose that there is a permanent increase in the rate of technological progress, so that g rises to g'. Sketch a graph of the growth rate of output per worker over time. Be sure to pay close attention to the transition dynamics.

5. *Can we save too much?* Consumption is equal to output minus investment: $c = (1 - s)y$. In the context of the Solow model with no technological progress, what is the savings rate that maximizes steady-state consumption per worker? What is the marginal product of capital in this steady state? Show this point in a Solow diagram. Be sure to draw the production function on the diagram, and show consumption and saving and a line indicating the marginal product of capital. Can we save too much?

6. *Solow (1956) versus Solow (1957).* In the Solow model with technological progress, consider an economy that begins in steady state with a rate of technological progress, g, of 2 percent. Suppose g rises permanently to 3 percent. Assume $\alpha = 1/3$.

   (a) What is the growth rate of output per worker before the change, and what happens to this growth rate in the long run?

   (b) Using equation (2.15), perform the growth accounting exercise for this economy, both before the change and after the economy has reached its new balanced growth path. (Hint: recall that $B \equiv A^{1-\alpha}$.) How much of the increase in the growth rate of output per worker is due to a change in the growth rate of capital per worker, and how much is due to a change in multifactor productivity growth?

   (c) In what sense does the growth accounting result in part (b) produce a misleading picture of this experiment?

3 EMPIRICAL APPLICATIONS OF NEOCLASSICAL GROWTH MODELS

This chapter considers several applications of the Solow model and its descendents, which we will group together under the rubric of "neoclassical growth models." In the first section of this chapter, we develop one of the key descendents of the Solow model, an extension that incorporates human capital. Then, we examine the "fit" of the model: How well does the neoclassical growth model explain why some countries are rich and others are poor? In the second section of this chapter, we examine the model's predictions concerning growth rates and discuss the presence or lack of "convergence" in the data. Finally, the third section of this chapter merges the discussion of the cross-country distribution of income levels with the convergence literature and examines the evolution of the world income distribution.

3.1 THE SOLOW MODEL WITH HUMAN CAPITAL

In an influential paper published in 1992, "A Contribution to the Empirics of Economic Growth," Gregory Mankiw, David Romer, and David Weil evaluated the empirical implications of the Solow model and concluded that it performed very well. They then noted that the "fit" of the model could be improved even more by extending the model to include human capital — that is, by recognizing that labor in different economies may possess different levels of education and different skills. Extending the Solow model to include human capital or skilled labor is relatively straightforward, as we shall see in this section.[1]

Suppose that output, Y, in an economy is produced by combining physical capital, K, with skilled labor, H, according to a constant-returns, Cobb-Douglas production function

$$Y = K^{\alpha}(AH)^{1-\alpha}, \tag{3.1}$$

where A represents labor-augmenting technology that grows exogenously at rate g.

Individuals in this economy accumulate human capital by spending time learning new skills instead of working. Let u denote the fraction of an individual's time spent learning skills, and let L denote the total amount of (raw) labor used in production in the economy.[2] We assume that unskilled labor learning skills for time u generates skilled labor H according to

$$H = e^{\psi u}L, \tag{3.2}$$

where ψ is a positive constant we will discuss in a moment. Notice that if $u = 0$, then $H = L$ — that is, all labor is unskilled. By increasing u, a unit of unskilled labor increases the *effective* units of skilled labor H. To see by how much, take logs and derivatives of equation (3.2) to see that

$$\frac{d\log H}{du} = \psi \Longrightarrow \frac{dH}{du} = \psi H. \tag{3.3}$$

To interpret this equation, suppose that u increases by 1 unit (think of this as one additional year of schooling), and suppose $\psi = .10$. In this

[1]The development here differs from that in Mankiw, Romer, and Weil (1992) in one important way. Mankiw, Romer, and Weil allow an economy to accumulate human capital in the same way that it accumulates physical capital: by foregoing consumption. Here, instead, we follow Lucas (1988) in assuming that individuals spend time accumulating skills, much like a student going to school. See Exercise 6 at the end of this chapter.

[2]Notice that if P denotes the total population of the economy, then the total amount of labor input in the economy is given by $L \equiv (1 - u)P$.

case, H rises by 10 percent. The fact that the effects are proportional is driven by the somewhat odd presence of the exponential e in the equation. This formulation is intended to match a large literature in labor economics that finds that an additional year of schooling increases the wages earned by an individual by something like 10 percent.[3]

Physical capital is accumulated by investing some output instead of consuming it, as in Chapter 2:

$$\dot{K} = s_K Y - dK, \tag{3.4}$$

where s_K is the investment rate for physical capital and d is the constant depreciation rate.

We solve this model using the same techniques employed in Chapter 2. First, we let lower-case letters denote variables divided by the stock of unskilled labor, L, and rewrite the production function in terms of output per worker as

$$y = k^{\alpha}(Ah)^{1-\alpha}. \tag{3.5}$$

Notice that $h = e^{\psi u}$. How do agents decide how much time to spend accumulating skills instead of working? Just as we assume that individuals save and invest a constant fraction of their income, we will assume that u is constant and given exogenously.[4]

The fact that h is constant means that the production function in equation (3.5) is very similar to that used in Chapter 2. In particular, along a balanced growth path, y and k will grow at the constant rate g, the rate of technological progress.

As in Chapter 2, the model is solved by considering "state variables" that are constant along a balanced growth path. There, recall that the state variables were terms such as y/A. Here, since h is constant, we can define the state variables by dividing by Ah. Denoting these state variables with a tilde, equation (3.5) implies that

$$\tilde{y} = \tilde{k}^{\alpha}, \tag{3.6}$$

which is the same as equation (2.11).

[3]Bils and Klenow (2000) apply this Mincerian formulation in the context of economic growth.

[4]We return to this issue in Chapter 7.

Following the reasoning from Chapter 2, the capital accumulation equation can be written in terms of the state variables as

$$\dot{\tilde{k}} = s_K \tilde{y} - (n + g + d)\tilde{k}. \tag{3.7}$$

Notice that in terms of state variables, this model is identical to the model we have already solved in Chapter 2. That is, equations (3.6) and (3.7) are identical to equations (2.11) and (2.12). This means that all of the results we discussed in Chapter 2 regarding the dynamics of the Solow model apply here. Adding human capital as we have done it does not change the basic flavor of the model.

The steady-state values of $\tilde{k}$ and $\tilde{y}$ are found by setting $\dot{\tilde{k}} = 0$, which yields

$$\frac{\tilde{k}}{\tilde{y}} = \frac{s_K}{n + g + d}.$$

Substituting this condition into the production function in equation (3.6), we find the steady-state value of the output-technology ratio $\tilde{y}$:

$$\tilde{y}^* = \left(\frac{s_K}{n + g + d}\right)^{\alpha/(1-\alpha)}.$$

Rewriting this in terms of output per worker, we get

$$y^*(t) = \left(\frac{s_K}{n + g + d}\right)^{\alpha/(1-\alpha)} hA(t), \tag{3.8}$$

where we have explicitly included t to remind us which variables are growing over time.

This last equation summarizes the explanation provided by the extended Solow model for why some countries are rich and others are poor. Countries are rich because they have high investment rates in physical capital, spend a large fraction of time accumulating skills ($h = e^{\psi u}$), have low population growth rates, and have high levels of technology. Furthermore, in the steady state, per capita output grows at the rate of technological progress, g, just as in the original Solow model.

How well does this model perform empirically in terms of explaining why some countries are richer than others? Because incomes are growing over time, it is useful to analyze the model in terms of *relative* incomes. If we define per capita income relative to the United States

to be

$$\hat{y}^* = \frac{y^*}{y^*_{US}},$$

then from equation (3.8), relative incomes are given by

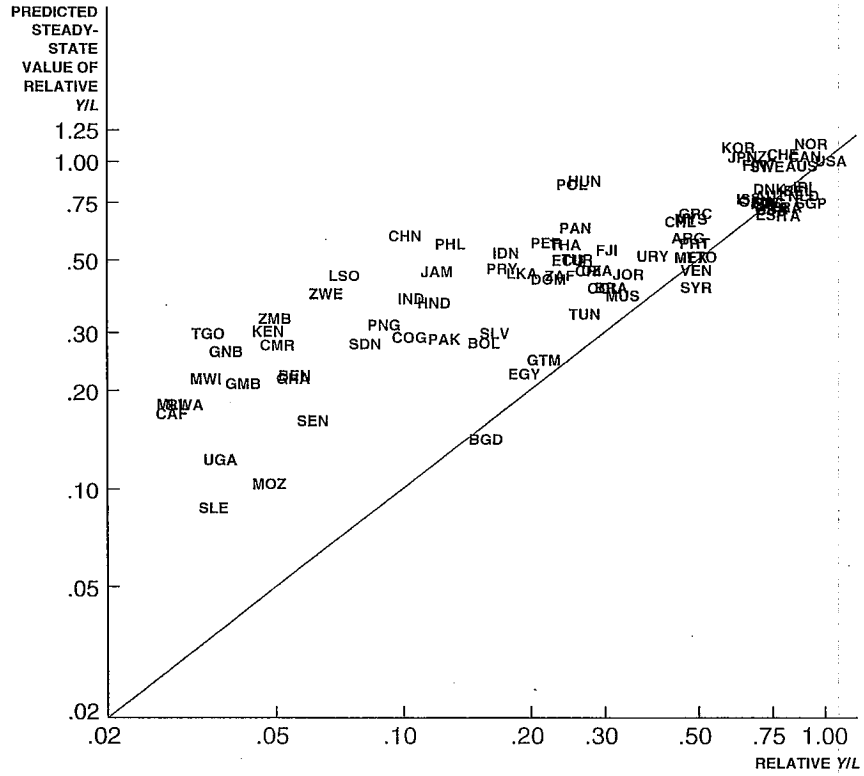
$$\hat{y}^* = \left(\frac{\hat{s}_K}{\hat{x}}\right)^{\alpha/(1-\alpha)} \hat{h}\hat{A}, \tag{3.9}$$

where the "hat" (^) is used to denote a variable relative to its U.S. value, and $x \equiv n+g+d$. Notice, however, that unless countries are all growing at the same rate, even relative incomes will not be constant. That is, if the United Kingdom and the United States are growing at different rates, then y_{UK}/y_{US} will not be constant.

In order for relative incomes to be constant in the steady state, we need to make the assumption that g is the same in all countries — that is, the rate of technological progress in all countries is identical. On the surface, this seems very much at odds with one of our key stylized facts from Chapter 1: that growth rates vary substantially across countries. We will discuss technology in much greater detail in later chapters, but for now, notice that if g varies across countries, then the "income gap" between countries eventually becomes infinite. This may not seem plausible if growth is driven purely by technology. Technologies may flow across international borders through international trade, or in scientific journals and newspapers, or through the immigration of scientists and engineers. It may be more plausible to think that technology transfer will keep even the poorest countries from falling too far behind, and one way to interpret this statement is that the growth rates of technology, g, are the same across countries. We will formalize this argument in Chapter 6. In the meantime, notice that in no way are we requiring the *levels* of technology to be the same; in fact, differences in technology presumably help to explain why some countries are richer than others.

Still, we are left wondering why it is that countries have grown at such different rates over the last thirty years if they have the same underlying growth rate for technology. It may seem that the Solow model cannot answer this question, but in fact it provides a very good answer that will be discussed in the next section. First, however, we return to the basic question of how well the extended Solow model fits the data.

FIGURE 3.1 THE "FIT" OF THE NEOCLASSICAL GROWTH MODEL, 1997



Note: A log scale is used for each axis.

By obtaining estimates of the variables and parameters in equation (3.9), we can examine the "fit" of the neoclassical growth model: empirically, how well does it explain why some countries are rich and others are poor?

Figure 3.1 compares the actual levels of GDP per worker in 1997 to the levels predicted by equation (3.9). To use the equation, we assume a physical capital share of $\alpha = 1/3$. This choice fits well with the observation that the share of GDP paid to capital is about 1/3. We measure u as the average educational attainment of the labor force (in years) and assume that $\psi = .10$. Such a value implies that each year of schooling

increases a worker's wage by 10 percent, a number roughly consistent with international evidence on returns to schooling.[5] In addition, we assume that $g + d = .075$ for all countries; we will discuss the assumption that g is the same in all countries in later chapters, and there is no good data on differences in d across countries. Finally, we assume that the technology level, A, is the same across countries. That is, we tie one hand behind our back to see how well the model performs without introducing technological differences. This assumption will be discussed shortly. The data used in this exercise are listed in Appendix C at the end of the book.

Without accounting for differences in technology, the neoclassical model still describes the distribution of per capita income across countries fairly well. Countries such as the United States and Norway are quite rich, as predicted by the model. Countries such as Uganda and Mozambique are decidedly poor. The main failure of the model — that it ignores differences in technology — can be seen by the departures from the 45-degree line in Figure 3.1: the model predicts that the poorest countries should be richer than they are.

How can we incorporate actual technology levels into the analysis? It is difficult to answer this question in a satisfactory manner, but there is a convenient "cheat" that is available. We can use the production function itself to solve for the level of A consistent with each country's output and capital. This is a cheat in that we are simply calculating A to make the model fit the data. However, it is an informative cheat. One can examine the As that are needed to fit the data to see if they are plausible.

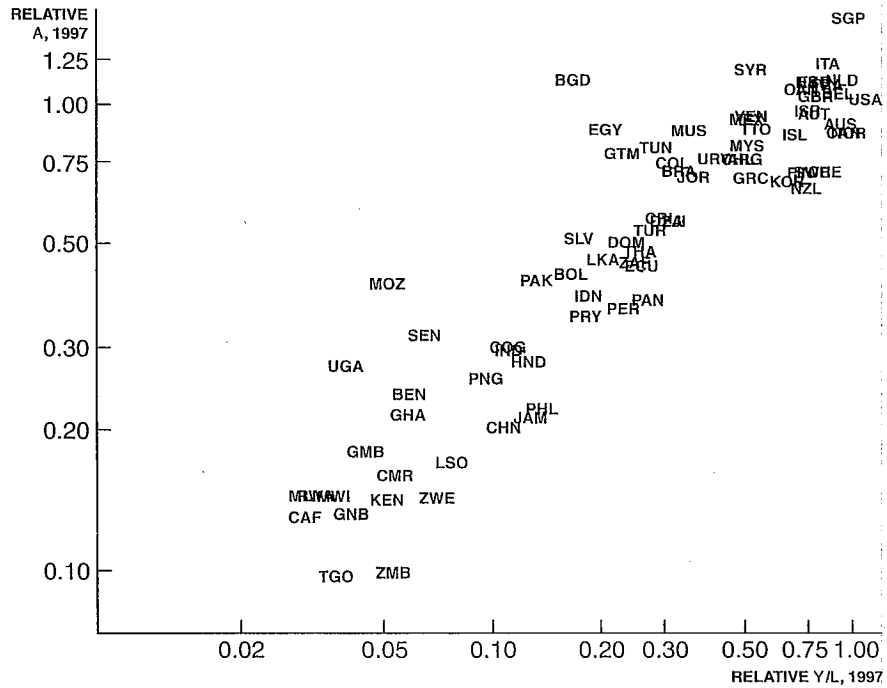
Solving the production function in Equation (3.5) for A yields

$$A = \left(\frac{y}{k}\right)^{\alpha/1-\alpha} \frac{y}{h}.$$

With data on GDP per worker, capital per worker, and educational attainment for each country, we can use this equation to estimate actual levels of A. These estimates are reported in Figure 3.2.

[5] See Jones (1996) for additional details. Notice that measuring u as years of schooling means that it is no longer between zero and one. This problem can be addressed by dividing years of schooling by potential lifespan, which simply changes the value of ψ proportionally and is therefore ignored.

FIGURE 3.2 **PRODUCTIVITY LEVELS, 1997**



Note: A log scale is used for each axis, and U.S. values are normalized to 1.

From this figure, one discovers several important things. First, the levels of A calculated from the production function are strongly correlated with the levels of output per worker across countries. Rich countries generally have high levels of A, and poor countries generally have low levels. Countries that are rich not only have high levels of physical and human capital, but they also manage to use these inputs very productively.

Second, although levels of A are highly correlated with levels of income, the correlation is far from perfect. Countries such as Singapore, Italy, and Bangladesh have much higher levels of A than would be expected from their GDP per worker, and perhaps have levels that are too high to be plausible. Indeed, several countries have levels of A higher than that in the United States. This observation leads to an important

remark. Estimates of A computed this way are like the residuals from growth accounting: they incorporate *any* differences in production not factored in through the inputs. For example, we have not controlled for differences in the quality of educational systems, the importance of experience at work and on-the-job training, or the general health of the labor force. These differences will therefore be included in A. In this sense, it is more appropriate to refer to these estimates as total factor productivity levels rather than technology levels.

Finally, the differences in total factor productivity across countries are large. The poorest countries of the world have levels of A that are only 10 to 15 percent of those in the richest countries.

With this observation, we can return to equation (3.9) to make one last remark. The richest countries of the world have an output per worker that is roughly 32 times that of the poorest countries of the world. This difference can be broken down into differences associated with investment rates in physical capital, investment rates in human capital, and differences in productivity. For this purpose, it is helpful to refer to the data in Appendix C. The richest countries of the world have investment rates that are around 25 percent, while the poorest countries of the world have investment rates around 5 percent. As a rough approximation, then, s/x varies by about a factor of 5 across countries. According to equation (3.9), it is the square root of this factor (since $\alpha/1 - \alpha = 1/2$) that contributes to output per worker, so that differences in physical capital account for just over a factor of 2 of the differences in output per worker between the rich and poor countries.

Similarly, workers in rich countries have about 10 or 11 years of education on average, whereas workers in poor countries have less than 3 years. Assuming a return to schooling of 10 percent, this suggests that $\hat{h} \approx e^{.10(11-3)} \approx e^{.8} \approx 2.2$. That is, differences in educational attainment also contribute a factor of just over 2 to differences in output per worker between the rich and poor countries.

What accounts for the remainder? By construction, differences in total factor productivity contribute the remaining factor of 8 to the differences in output per worker between the rich and poor countries.[6]

[6] A more extensive analysis of productivity levels can be found in Klenow and Rodriguez-Clare (1997) and Hall and Jones (1999).

Productivity differences across countries are large, and a satisfactory theory of growth and development needs to explain these differences.

In summary, the Solow framework is extremely successful in helping us to understand the wide variation in the wealth of nations. Countries that invest a large fraction of their resources in physical capital and in the accumulation of skills are rich. Countries that use these inputs productively are rich. The countries that fail in one or more of these dimensions suffer a corresponding reduction in income. Of course, one thing the Solow model does not help us understand is *why* some countries invest more than others, and *why* some countries attain higher levels of technology or productivity. Addressing these questions is the subject of Chapter 7. As a preview, the answers are tied intimately to government policies and institutions.

3.2 CONVERGENCE AND EXPLAINING DIFFERENCES IN GROWTH RATES

We have discussed in detail the ability of the neoclassical model to explain differences in income levels across economies, but how well does it perform at explaining differences in growth rates? An early hypothesis proposed by economic historians such as Aleksander Gerschenkron (1952) and Moses Abramovitz (1986) was that, at least under certain circumstances, "backward" countries would tend to grow faster than rich countries, in order to close the gap between the two groups. This catch-up phenomenon is referred to as *convergence*. For obvious reasons, questions about convergence have been at the heart of much empirical work on growth. We documented in Chapter 1 the enormous differences in levels of income per person around the world: the typical person in the United States earns in less than ten days the annual income of the typical person in Ethiopia. The question of convergence asks whether these enormous differences are getting smaller over time.

An important cause of convergence might be technology transfer, but the neoclassical growth model provides another explanation for convergence that we will explore in this section. First, however, let's examine the empirical evidence on convergence.

William Baumol (1986), alert to the analysis provided by economic historians, was one of the first economists to provide statistical evidence documenting convergence among some countries and the absence of convergence among others. The first piece of evidence presented by Baumol is displayed in Figure 3.3, which plots per capita GDP (on a log scale) for several industrialized economies from 1870 to 1994. The narrowing of the gaps between countries is evident in this figure. Interestingly, the world "leader" in terms of per capita GDP in 1870 was Australia (not shown). The United Kingdom had the second-highest per capita GDP and was recognized as the industrial center of the Western world. Around the turn of the century, the United States surpassed Australia and the United Kingdom and has remained the "leader" ever since.

FIGURE 3.3 **PER CAPITA GDP, 1870–1994**

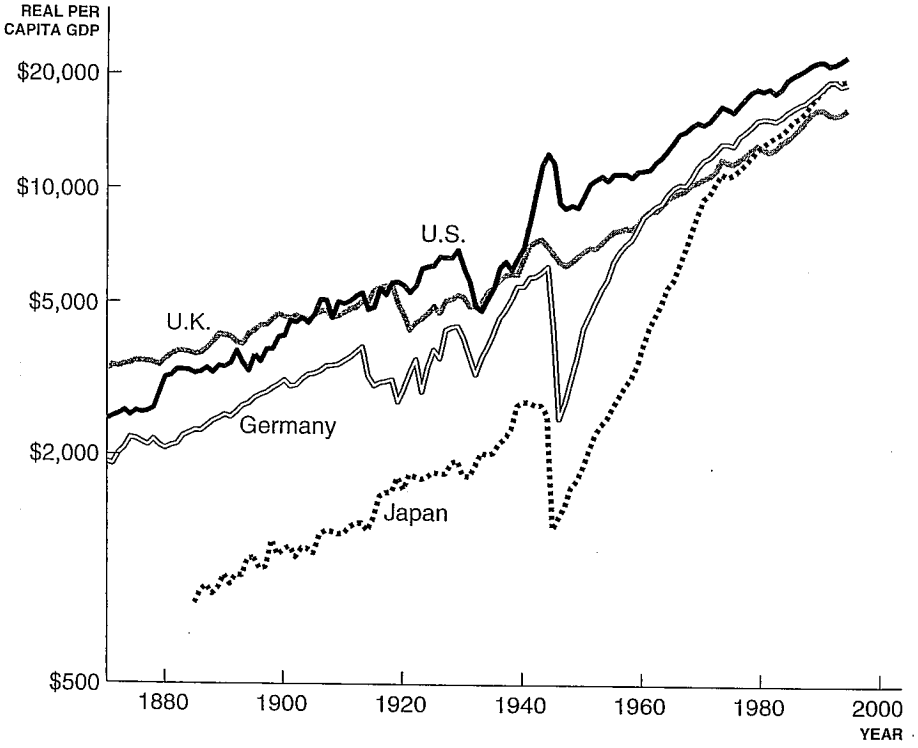


FIGURE 3.4 **GROWTH RATE VERSUS INITIAL PER CAPITA GDP, 1885–1994**

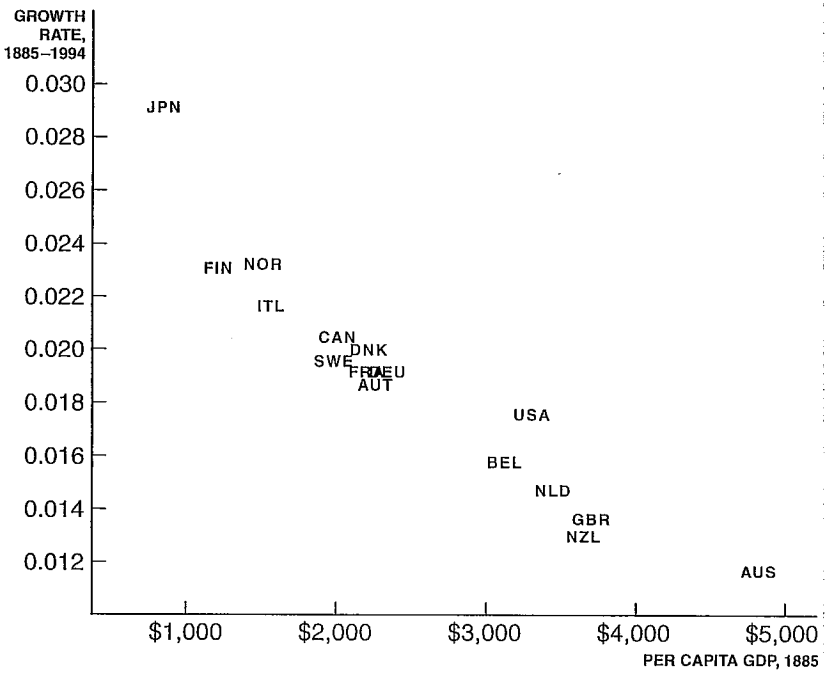
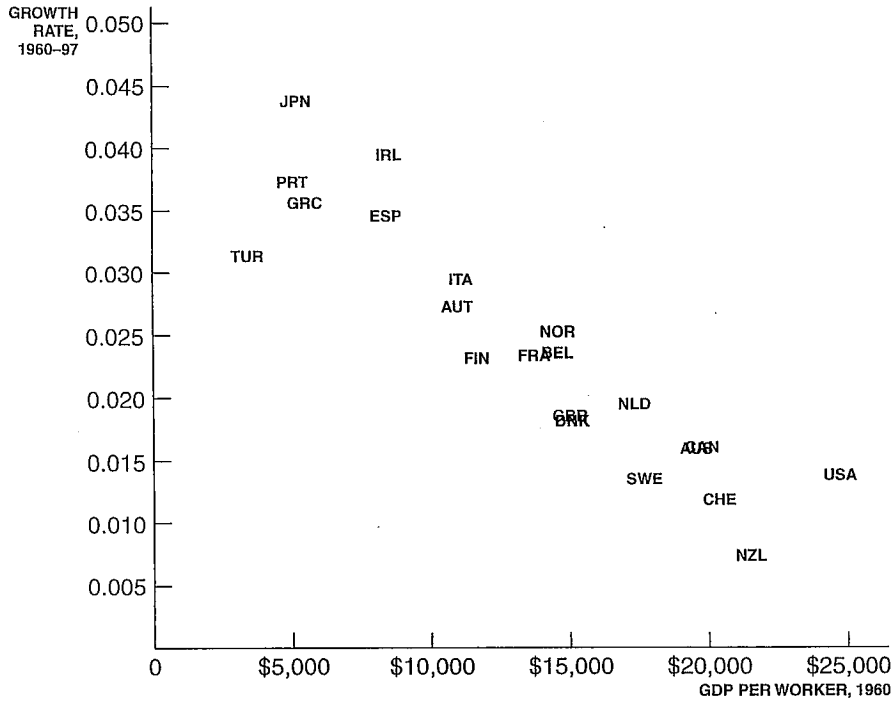


Figure 3.4 reveals the ability of the convergence hypothesis to explain why some countries grew fast and others grew slowly over the course of the last century. The graph plots a country's initial per capita GDP (in 1885) against the country's growth rate from 1885 to 1994. The figure reveals a strong negative relationship between the two variables: countries such as Australia and the United Kingdom, which were relatively rich in 1885, grew most slowly, while countries like Japan that were relatively poor grew most rapidly. The simple convergence hypothesis seems to do a good job of explaining differences in growth rates, at least among this sample of industrialized economies.[7]

Figures 3.5 and 3.6 plot growth rates versus initial GDP per worker for the countries that are members of the Organization for Economic

[7] J. Bradford DeLong (1988) provides an important criticism of this result. See Exercise 5 at the end of this chapter.

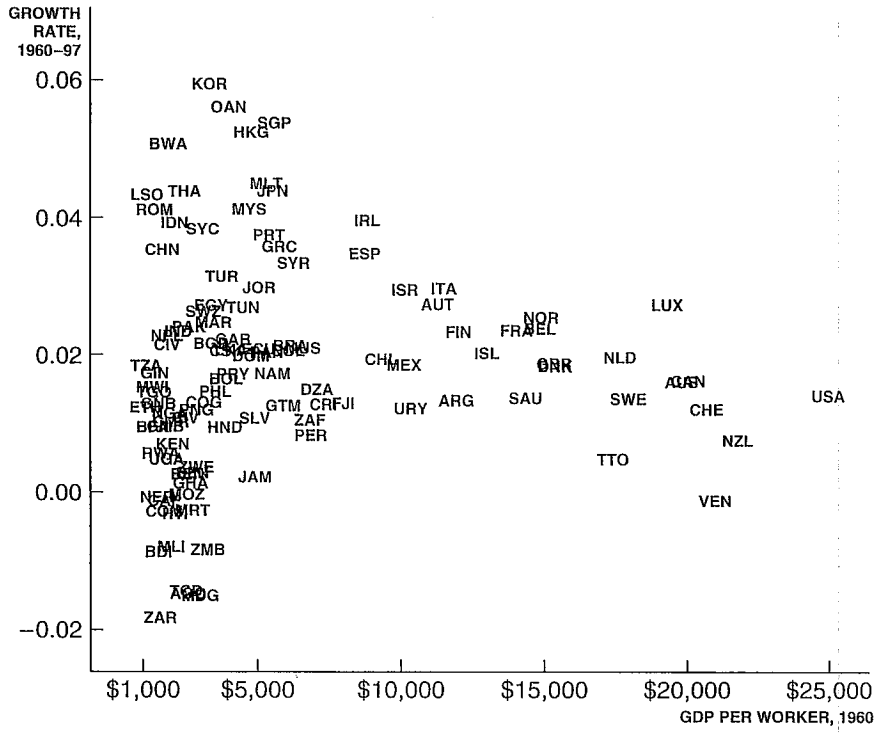
FIGURE 3.5 **CONVERGENCE IN THE OECD, 1960–97**



Cooperation and Development (OECD) and for the world for the period 1960–97. Figure 3.5 shows that the convergence hypothesis works extremely well for explaining growth rates across the OECD for the period examined. But before we declare the hypothesis a success, note that Figure 3.6 shows that the convergence hypothesis fails to explain differences in growth rates across the world as a whole. Baumol also reported this finding: across large samples of countries, it does not appear that poor countries grow faster than rich countries. The poor countries are not "closing the gap" that exists in per capita incomes. (Recall that Table 1.1 in Chapter 1 supports this finding.)

Why, then, do we see convergence among some sets of countries but a lack of convergence among the countries of the world as a whole? The neoclassical growth model suggests an important explanation for these findings.

FIGURE 3.6 **THE LACK OF CONVERGENCE FOR THE WORLD, 1960–97**



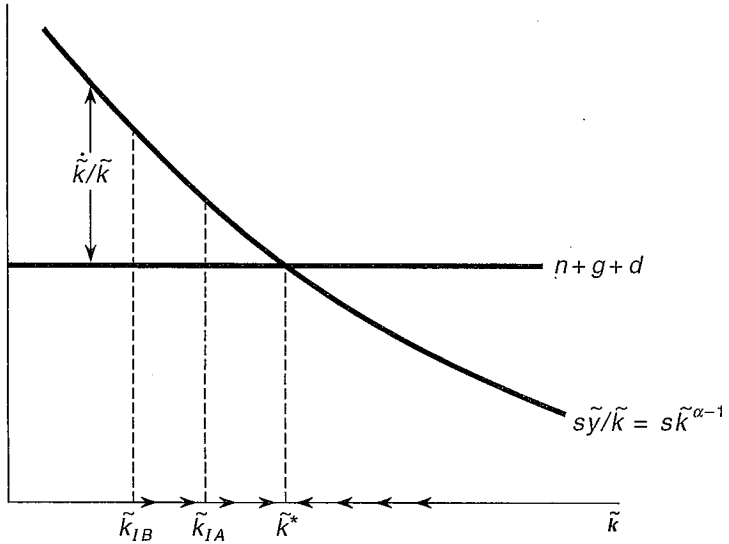
Consider the key differential equation of the neoclassical growth model, given in equation (3.7). This equation can be rewritten as

$$\frac{\dot{\tilde{k}}}{\tilde{k}} = s_K \frac{\tilde{y}}{\tilde{k}} - (n + g + d). \tag{3.10}$$

Remember that $\tilde{y}$ is equal to $\tilde{k}^{\alpha}$. Therefore, the average product of capital $\tilde{y}/\tilde{k}$ is equal to $\tilde{k}^{\alpha-1}$. In particular, it declines as $\tilde{k}$ rises, because of the diminishing returns to capital accumulation in the neoclassical model.

As in Chapter 2, we can analyze this equation in a simple diagram, shown in Figure 3.7. The two curves in the figure plot the two terms on the right-hand side of equation (3.10). Therefore, the difference between the curves is the growth rate of $\tilde{k}$. Notice that the growth rate of $\tilde{y}$ is simply proportional to this difference. Furthermore, because the growth

FIGURE 3.7 **TRANSITION DYNAMICS IN THE NEOCLASSICAL MODEL**



rate of technology is constant, any changes in the growth rates of $\tilde{k}$ and $\tilde{y}$ must be due to changes in the growth rates of capital per worker, k, and output per worker, y.

Suppose the economy of InitiallyBehind starts with the capital-technology ratio $\tilde{k}_{IB}$ shown on Figure 3.7, while the neighboring economy of InitiallyAhead starts with the higher capital-technology ratio indicated by $\tilde{k}_{IA}$. If these two economies have the same levels of technology, the same rates of investment, and the same rates of population growth, then InitiallyBehind will temporarily grow faster than InitiallyAhead. The output-per-worker gap between the two countries will narrow over time as both economies approach the same steady state. An important prediction of the neoclassical model is this: *Among countries that have the same steady state, the convergence hypothesis should hold: poor countries should grow faster on average than rich countries.*

For the industrialized countries, the assumption that their economies have similar technology levels, investment rates, and population

growth rates may not be a bad one. The neoclassical model, then, would predict the convergence that we saw in Figures 3.4 and 3.5. This same reasoning suggests a compelling explanation for the *lack* of convergence across the world as a whole: all countries do not have the same steady states. In fact, as we saw in Figure 3.2, the differences in income levels around the world largely reflect differences in steady states. Because all countries do not have the same investment rates, population growth rates, or technology levels, they are not generally expected to grow toward the same steady-state target.

Another important prediction of the neoclassical model is related to growth rates. This prediction, which can be found in many growth models, is important enough that we will give it a name, the "principle of transition dynamics":

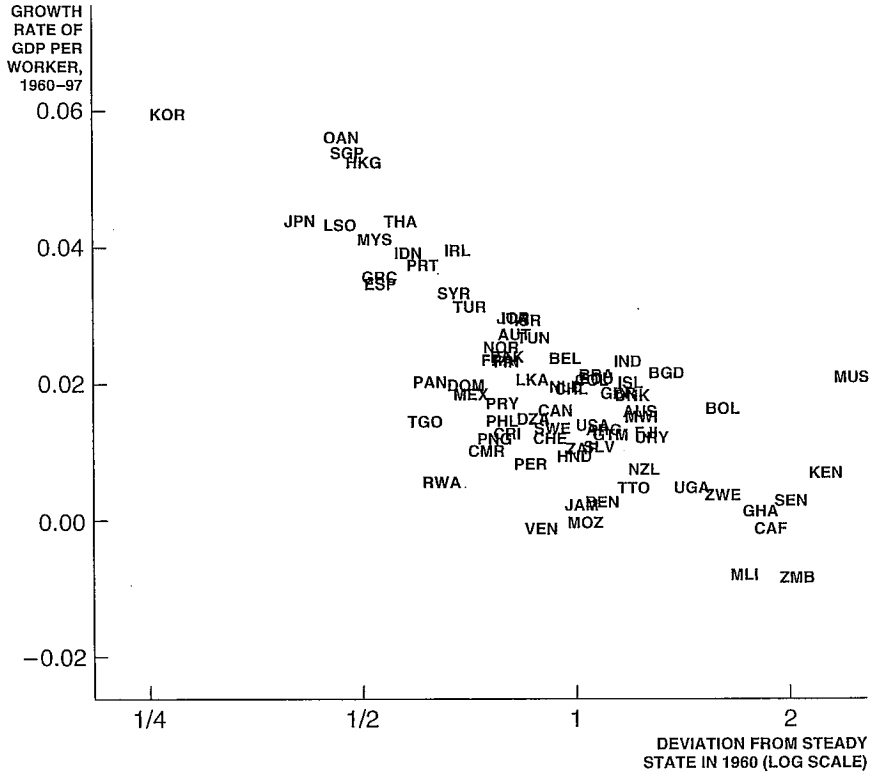
> **The further an economy is "below" its steady state, the faster the economy should grow. The further an economy is "above" its steady state, the slower the economy should grow.**[8]

This principle is clearly illustrated by the analysis of equation (3.10) provided in Figure 3.7. Although it is a key feature of the neoclassical model, the principle of transition dynamics applies much more broadly. In Chapters 5 and 6, for example, we will see that it is also a feature of the models of new growth theory that endogenize technological progress.

Mankiw, Romer, and Weil (1992) and Barro and Sala-i-Martin (1992) show that this prediction of the neoclassical model can explain differences in growth rates across the countries of the world. Figure 3.8 illustrates this point by plotting the growth rate of GDP per worker from 1960 to 1997 against the deviation of GDP per worker (relative to that of the U.S.) from its steady-state value. This steady state is computed according to equation (3.9) using the data in Appendix C and a total factor productivity level from 1970. (You will be asked to undertake a similar calculation in Exercise 1 at the end of the chapter.) Comparing Figure 3.6 and Figure 3.8, one sees that although poorer countries do not necessarily grow faster, countries that are "poor" relative to their own steady states do tend to grow more rapidly. In 1960, good

[8]In simple models, including most of those presented in this book, this principle works well. In more complicated models with more state variables, however, it must be modified.

FIGURE 3.8 "CONDITIONAL" CONVERGENCE FOR THE WORLD, 1960–97



Note: The variable on the x-axis is $\hat{y}_{60}/\hat{y}^*$. Estimates of A for 1970 are used to compute the steady state.

examples of these countries were Korea, Japan, Singapore, and Hong Kong — economies that grew rapidly over the next forty years, just as the neoclassical model would predict.[9]

[9]Mankiw, Romer, and Weil (1992) and Barro and Sala-i-Martin (1992) have called this phenomenon "conditional convergence," because it reflects the convergence of countries after we control for ("condition on") differences in steady states. It is important to keep in mind what this "conditional convergence" result means. It is simply a confirmation of a result predicted by the neoclassical growth model: that countries with similar steady states will exhibit convergence. It does not mean that all countries in the world are converging to the same steady state, only that they are converging to their own steady states according to a common theoretical model.

This analysis of convergence has been extended by a number of authors to different sets of economies. For example, Barro and Sala-i-Martin (1991, 1992) show that the U.S. states, regions of France, and prefectures in Japan all exhibit "unconditional" convergence similar to what we've observed in the OECD. This matches the prediction of the Solow model if regions within a country are similar in terms of investment and population growth, as seems reasonable.

How does the neoclassical model account for the wide differences in growth rates across countries documented in Chapter 1? The principle of transition dynamics provides the answer: countries that have not reached their steady states are not expected to grow at the same rate. Those "below" their steady states will grow rapidly, and those "above" their steady states will grow slowly.

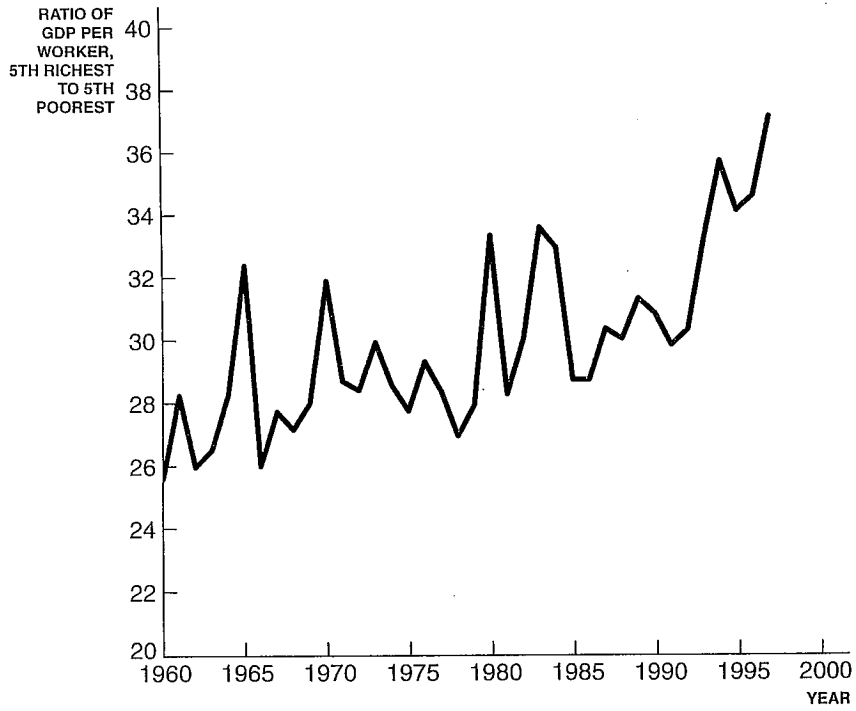
As we saw in Chapter 2, there are many reasons why countries may not be in steady state. An increase in the investment rate, a change in the population growth rate, or an event like World War II that destroys much of a country's capital stock will generate a gap between current income and steady-state income. This gap will change growth rates until the economy returns to its steady-state path. Other "shocks" can also cause temporary differences in growth rates. For example, large changes in oil prices will have important effects on the economic performance of oil-exporting countries. Mismanagement of the macroeconomy can similarly generate temporary changes in growth performance. The hyperinflations in many Latin American countries during the 1980s are a good example of this. Working in the other direction, policy reforms that shift the steady-state path of an economy upward can generate increases in growth rates along a transition path. Increases in the investment rate, skill accumulation, or the level of technology will have this effect.[10]

3.3 THE EVOLUTION OF THE INCOME DISTRIBUTION

Convergence, the closing of the gap between rich and poor economies, is just one possible outcome among many that could be occurring. Alternatively, perhaps the poorest countries are falling behind while countries

[10]Barro (1991) and Easterly, Kremer, et al. (1993) provide empirical analyses of why countries have exhibited different growth rates since 1960.

FIGURE 3.9 **INCOME RATIOS, 5TH-RICHEST COUNTRY TO 5TH-POOREST COUNTRY, 1960–97**



with "intermediate" incomes are converging toward the rich. Or perhaps countries are not getting any closer together at all but are instead fanning out, with the rich countries getting richer and the poor countries getting poorer. More generally, these questions are really about the evolution of the distribution of per capita incomes around the world.[11]

Figure 3.9 illustrates a key fact about the evolution of the income distribution: for the world as a whole, the enormous gaps in income across countries have generally not narrowed over time. This figure

[11]Jones (1997) provides an overview of the literature on the world income distribution. Quah (1993, 1996) discusses this topic in more detail.

plots the ratio of GDP per worker in the 5th-richest country to GDP per worker in the 5th-poorest country. In 1960, GDP per worker of the fifth-richest country was more than 25 times that of the fifth-poorest country. By 1990, the ratio had risen slightly, to around 30. The 1990s witnessed an even sharper increase, to more than 35 by the end of the sample.

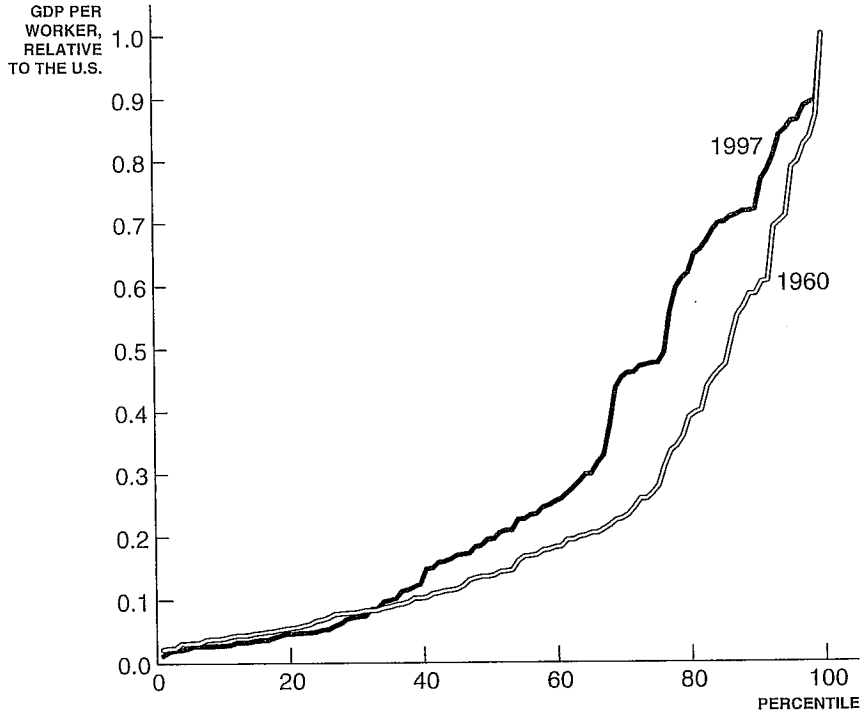
The widening of the world income distribution is a fact that almost certainly characterizes the world economy over its entire history. Incomes cannot get much lower than about $250: below this level widespread starvation and death set in. This number provides a lower bound on incomes at any date in the past, and this lower bound comes close to being attained by the poorest countries in the world even today. On the other hand, the incomes of the richest countries have been growing over time. This suggests that the ratio of the incomes in the richest to those in the poorest countries has also been rising. Lant Pritchett (1997), in a paper titled "Divergence, Big Time," calculates that the ratio of per capita GDP between the richest and poorest countries in the world was only 8.7 in 1870 but rose to 45.2 by 1990. Before 1870, the ratio was presumably even lower.

Whether this widening will continue in the future is an open question. One possible explanation for the increase is that countries climb onto the modern economic growth "escalator" at different points in time. As long as there are some countries that have yet to get on, the world income distribution widens. Once all countries get on, however, this widening may reverse.[12]

While Figure 3.9 shows that the "width" of the income distribution has increased, Figure 3.10 examines changes at each point in the income distribution. According to the figure, 50 percent of the countries had relative incomes that were less than 15 percent of U.S. GDP per worker in 1960; 80 percent of the countries had relative incomes less than 40 percent of U.S. GDP per worker. By 1997, these numbers had improved, particularly at the upper end: the 50th percentile was slightly less than 20 percent of U.S. GDP per worker while the 80th percentile was more than 60 percent. In contrast, the poorest economies — those below the 30th percentile, for example — actually had relative incomes in 1997 lower than in 1960. In this sense, one might say there was some catch-up or convergence at the middle and top of the income distribu-

[12]Robert E. Lucas, Jr. (2000), analyzes a model like this in a very readable manner.

FIGURE 3.10 THE EVOLUTION OF THE WORLD INCOME DISTRIBUTION



Note: A point (x, y) in the figure indicates that x percent of countries had relative GDP per worker less than or equal to y. One hundred ten countries are represented.

tion from 1960 to 1997, but divergence at the bottom end.[13] Danny Quah (1996) suggests that this tendency for the middle-income countries to become relatively richer while the poorest countries become relatively (but not necessarily absolutely) poorer will result in an income distribution with "twin peaks"—i.e., a mass of countries at both ends of the income distribution.

[13]It is interesting to compare this figure to the results in Chapter 1. An important difference is that the unit of observation here is the *country*; the unit of observation for the distributions computed in Chapter 1 was the *individual*.

EXERCISES

1. *Where are these economies headed?* Consider the following data:

	$\hat{y}_{97}$	s_K	u	n	$\hat{A}_{90}$
U.S.A.	1.000	0.204	11.9	0.010	1.000
Canada	0.864	0.246	11.4	0.012	0.972
Argentina	0.453	0.144	8.5	0.014	0.517
Thailand	0.233	0.213	6.1	0.015	0.468
Cameroon	0.048	0.102	3.4	0.028	0.234

Assume that $g + d = .075$, $\alpha = 1/3$, and $\psi = .10$ for all countries. Using equation (3.9), estimate the steady-state incomes of these economies, relative to the United States. Consider two extreme cases: (a) the 1990 TFP ratios are maintained, and (b) the TFP levels converge completely. For each case, which economy will grow fastest in the next decade and which slowest? Why?

2. *Policy reforms and growth.* Suppose an economy, starting from an initial steady state, undertakes new policy reforms that raise its steady-state level of output per worker. For each of the following cases, calculate the proportion by which steady-state output per worker increases and, using the slope of the relationship shown in Figure 3.8, make a guess as to the amount by which the growth rate of GDP per worker will be higher during the next forty years. Assume $\alpha = 1/3$ and $\psi = .10$. **(a)** The level of total factor productivity, A is permanently doubled. **(b)** The investment rate, s_K, is permanently doubled. **(c)** The average educational attainment of the labor force, u, is permanently increased by 5 years.

3. *What are state variables?* The basic idea of solving dynamic models that contain a differential equation is to first write the model so that along a balanced growth path, some state variable is constant. In Chapter 2, we used y/A and k/A as state variables. In this chapter, we used y/Ah and k/Ah. Recall, however, that h is a constant. This reasoning suggests that one should be able to solve the model using y/A and k/A as the state variables. Do this. That is, solve the growth model in equations (3.1) to (3.4) to get the solution in equation (3.8) using y/A and k/A as state variables.

4. *Galton's fallacy* (based on Quah 1993). During the late 1800s, Sir Francis Galton, a famous statistician in England, studied the distribution of heights in the British population and how the distribution was evolving over time. In particular, Galton noticed that the sons of tall fathers tended to be shorter than their fathers, and vice versa. Galton worried that this implied some kind of regression toward "mediocrity."

   Suppose that we have a population of 10 mothers who have 10 daughters. Suppose that their heights are determined as follows. Place 10 sheets of paper in a hat labeled with heights of 5′1″, 5′2″, 5′3″, ... 5′10″. Draw a number from the hat and let that be the height for a mother. Without replacing the sheet just drawn, continue. Now suppose that the heights of the daughters are determined in the same way, starting with the hat full again and drawing new heights. Make a graph of the change in height between daughter and mother against the height of the mother. Will tall mothers tend to have shorter daughters, and vice versa?

   Let the heights correspond to income levels, and consider observing income levels at two points in time, say 1960 and 1990. What does Galton's fallacy imply about a plot of growth rates against initial income? Does this mean the figures in this chapter are useless?[14]

5. *Reconsidering the Baumol results.* J. Bradford DeLong (1988), in a comment on Baumol's convergence result for the industrialized countries over the last century, pointed out that the result could be driven by the procedure through which the countries were selected. In particular, DeLong noted two things. First, only countries that were rich at the end of the sample (i.e., in the 1980s) were included. Second, several countries not included, such as Argentina, were richer than Japan in 1870. Use these points to criticize and discuss the Baumol results. Do these criticisms apply to the results for the OECD? For the world?

6. *The Mankiw-Romer-Weil (1992) model.* As mentioned in this chapter, the extended Solow model that we have considered differs slightly from that in Mankiw, Romer, and Weil (1992). This problem asks you to solve their model. The key difference is the treatment of human capital. Mankiw, Romer, and Weil assume that human capital is accumulated just like physical capital, so that it is measured in units of output instead of years of time.

   Assume production is given by $Y = K^{\alpha}H^{\beta}(AL)^{1-\alpha-\beta}$, where α and β are constants between zero and one whose sum is also between zero and one. Human capital is accumulated just like physical capital:

$$\dot{H} = s_H Y - dH,$$

   where s_H is the constant share of output invested in human capital. Assume that physical capital is accumulated as in equation (3.4), that the labor force grows at rate n, and that technological progress occurs at rate g. Solve the model for the path of output per worker $y \equiv Y/L$ along the balanced growth path as a function of s_K, s_H, n, g, d, α, and β. Discuss how the solution differs from that in equation (3.8). (Hint: Define state variables such as y/A, h/A, and k/A.)

[14] See Quah (1993) and Friedman (1992).

4 THE ECONOMICS OF IDEAS

The neoclassical models we have studied so far are in many ways capital-based theories of economic growth. These theories focus on modeling the accumulation of physical and human capital. In another sense, however, the theories emphasize the importance of technology. For example, the models do not generate economic growth in the absence of technological progress, and productivity differences help to explain why some countries are rich and others are poor. In this way, neoclassical growth theory highlights its own shortcoming: although technology is a central component of neoclassical theory, it is left unmodeled. Technological improvements arrive exogenously at a constant rate, g, and differences in technologies across economies are unexplained. In this chapter, we will explore the broad issues associated with creating an economic model of technology and technological improvement.

4.1 WHAT IS TECHNOLOGY?

In the economics of growth and development, the term technology has a very specific meaning: *technology* is the way inputs to the production process are transformed into output. For example, if we have a general production function $Y = F(K, L, \cdot)$, then the technology of production is given by the function $F(\cdot)$; this production function explains how inputs are transformed into output. In the Cobb-Douglas production function of earlier chapters, $Y = K^{\alpha}(AL)^{1-\alpha}$, A is an index of technology.[1]

Ideas improve the technology of production. A new idea allows a given bundle of inputs to produce more or better output. A good example of an idea was provided by Paul Romer (1990). Neanderthals used iron oxide as a pigment to create drawings on the walls of caves. Now, we "paint" iron oxide onto magnetic tape to produce VCR recordings. The "idea" behind the VCR allows us to use a given bundle of inputs to produce output that generates a higher level of utility. In the context of the production function above, a new idea generates an increase in the technology index, A.

Examples of ideas and technological improvements abound. Moore's Law (attributed to the former chairman of Intel, Gordon Moore) asserts that the number of transistors that can be packed onto a computer chip doubles approximately every 18 months. In 1800, light was provided by candles and oil lamps, whereas today we have very efficient fluorescent bulbs. William Nordhaus (1994) has calculated that the quality-adjusted price of light has fallen by a factor of 4,000 since the year 1800.[2]

Ideas are by no means limited to feats of engineering, however. Sam Walton's creation of the Wal-Mart approach to retailing is no less an idea than advances in semiconductor technology. The multiplex theater and diet soft drinks are innovations that allowed firms to combine inputs in new ways that consumers, according to revealed preference, have found very valuable. The assembly lines and mass production techniques that allowed Henry Ford's company to turn out a Model T every 24 seconds, and Ford's payment of wages of $5 per day when the prevailing wage was less than half that amount are business innovations that profoundly changed U.S. manufacturing.

[1] The parameter α is also part of the "technology" of production.
[2] See the *Economist*, October 22, 1994, p. 84.

THE ECONOMICS OF IDEAS

Beginning in the mid-1980s, Paul Romer formalized the relationship between the economics of ideas and economic growth.[3] This relationship can be thought of in the following way:

Ideas ⟶ Nonrivalry ⟶ Increasing Returns ⟶ Imperfect Competition

According to Romer, an inherent characteristic of ideas is that they are nonrivalrous. This nonrivalry implies the presence of increasing returns to scale. And to model these increasing returns in a competitive environment with intentional research necessarily requires imperfect competition. Each of these terms and the links between them will now be discussed in detail. In the next chapter, we will develop the mathematical model that integrates this reasoning.

A crucial observation emphasized by Romer (1990) is that ideas are very different from most other economic goods. Most goods, such as compact disc (CD) players or lawyer services are *rivalrous*. That is, my use of a CD player excludes your use of the same CD player, or my seeing a particular attorney today from 1:00 P.M. to 2:00 P.M. precludes your seeing the same attorney at the same time. Most economic goods share this property: the use of the good by one person precludes its use by another. If one thousand people each want to use a CD player, we have to provide them with one thousand CD players.

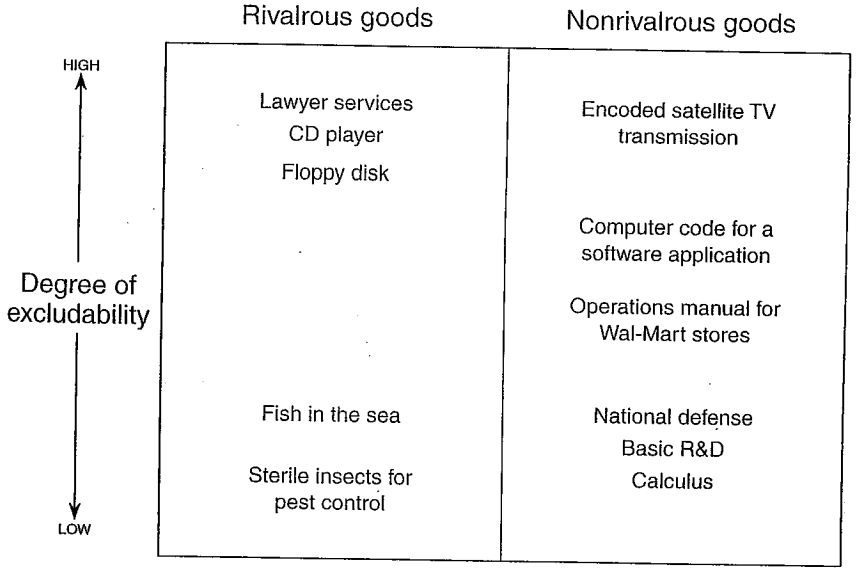
In contrast, ideas are *nonrivalrous*. The fact that Toyota takes advantage of just-in-time inventory methods does not preclude GM from taking advantage of the same technique. Once an idea has been created, anyone with knowledge of the idea can take advantage of it. Consider the design for the next-generation computer chip. Once the design itself has been created, factories throughout the country and even the world can use the design simultaneously to produce computer chips, provided they have the plans in hand. The paper the plans are written on is rivalrous; an engineer, whose skills are needed to understand the plans, is rivalrous; but the instructions written on the paper — the ideas — are not.

[3]This basic insight is found in Shell (1967), Phelps (1968), Nordhaus (1969), and Romer (1986).

This last observation suggests another important characteristic of ideas, one that ideas share with most economic goods: they are, at least partially, *excludable*. The degree to which a good is excludable is the degree to which the owner of the good can charge a fee for its use. The firm that invents the design for the next computer chip can presumably lock the plans in a safe and restrict access to the design, at least for some period of time. Alternatively, copyright and patent systems grant inventors who receive copyrights or patents the right to charge for the use of their ideas.

Figure 4.1, taken in large part from Romer (1993), lists a variety of economic goods according to their degree of excludability and whether they are rivalrous or nonrivalrous. Both rivalrous and nonrivalrous goods vary in the degree to which they are excludable. Goods such as a CD player, a floppy disk, or the services of a lawyer are highly excludable.

FIGURE 4.1 **ECONOMIC ATTRIBUTES OF SELECTED GOODS**



SOURCE: This is a slightly altered version of Figure 1 in Romer (1993).

Goods that suffer from the "tragedy of the commons" problem are rivalrous but have a low degree of excludability.[4] The classic example of such goods is the overgrazing of common land shared by English peasants during the Middle Ages. The cost of one peasant's choosing to graze an additional cow on the commons is shared by all of the peasants, but the benefit is captured solely by one peasant. The result is an inefficiently high level of grazing that can potentially destroy the commons. A similar outcome occurs when a group of friends goes to a nice restaurant and divides the bill evenly at the end of the evening—suddenly everyone wants to order an expensive bottle of wine and a rich chocolate dessert. A modern example of the commons problem is the overfishing of international waters.

Ideas are nonrivalrous goods, but they vary substantially in their degree of excludability. Encoded satellite TV transmissions are highly excludable, whereas computer software is less excludable. Both of these goods or ideas are essentially a collection of 1's and 0's ordered in a particular way so as to convey information. The digital signals of an encoded satellite transmission are scrambled so as to be useful only to someone with a decoder. In contrast, computer software is often "unscrambled": anyone with a disk drive can copy software to give to a friend. Software companies take advantage of this aspect of ideas in manufacturing software but can also find it to be a problem because of software pirating. Similar considerations apply to the operating manual for Wal-Mart. Sam Walton details his ideas for efficiently running a retail operation in the manual and gives it to all of his stores. However, some of these ideas may be copied by an astute observer of Wal-Mart's business behavior.

Nonrivalrous goods that are essentially unexcludable are often called *public goods*. A traditional example is national defense. For example, consider the often-debated "Star Wars" defense shield that would protect the United States from hostile missiles. If the shield is going to protect *some* citizens in Washington, D.C., it will protect *all* citizens in the nation's capital; the "Star Wars" defense system is nonrivalrous and unexcludable. Some ideas may also be both nonrivalrous and unexcludable. For example, the results of basic research and development (R&D) may by their very nature be unexcludable. Calculus, our scientific understanding of medicine, and the Black-Scholes formula for pricing financial options are other examples.[5]

[4] See Hardin (1968).

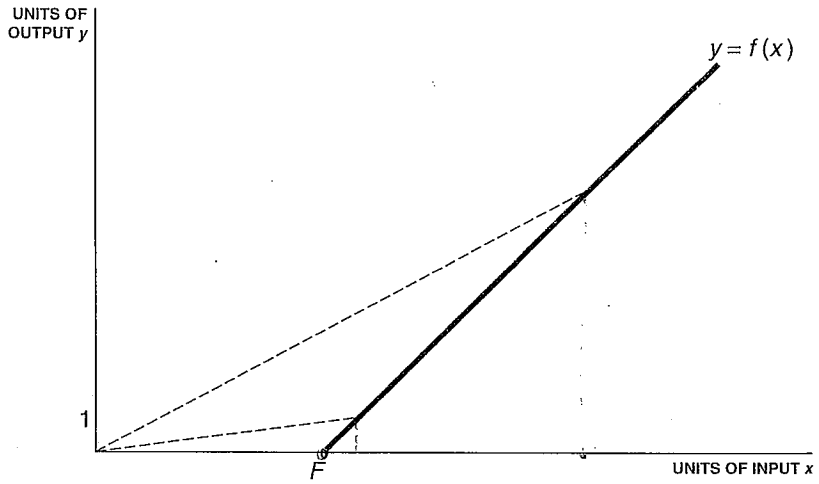
The economics of goods depends on their attributes. Goods that are excludable allow their producers to capture the benefits they produce; goods that are not excludable involve substantial "spillovers" of benefits that are not captured by producers. Such spillovers are called *externalities*. Goods with positive spillovers tend to be underproduced by markets, providing a classic opportunity for government intervention to improve welfare. For example, basic R&D and national defense are financed primarily by the government. Goods with negative spillovers may be overproduced by markets, and government regulation may be needed if property rights cannot be well defined. The tragedy of the commons is a good example.

Goods that are rivalrous must be produced each time they are sold; goods that are nonrivalrous need be produced only once. That is, nonrivalrous goods such as ideas involve a fixed cost of production and zero marginal cost. For example, it costs a great deal to produce the first unit of the latest word processor or spreadsheet, but subsequent units are produced simply by copying the software from the first unit. It required a great deal of inspiration and perspiration for Thomas Edison and his lab to produce the first commercially viable electric light. But once the first light was produced, additional lights could be produced at a much lower per-unit cost. In both the spreadsheet and the lightbulb examples, notice that the only reason for a nonzero marginal cost is that the nonrivalrous good—the idea—is embodied in a rivalrous good—the CD or the materials of the lightbulb.

This reasoning leads to a simple but powerful insight: the economics of "ideas" is intimately tied to the presence of increasing returns to scale and imperfect competition. The link to increasing returns is almost immediate once we grant that ideas are associated with fixed costs. Returning to the software example, the "idea" underlying the next generation of word processing (perhaps with voice recognition, let's say)

[5] Fischer Black and Myron Scholes (1972) developed an elegant mathematical technique for pricing a financial security called an option. The formula, the basis for the 1997 Nobel Prize in Economics, is widely used on Wall Street and throughout the financial community.

FIGURE 4.2 FIXED COSTS AND INCREASING RETURNS



requires a one-time research cost. Once the product is developed, each additional unit is produced with constant returns to scale: doubling the number of CDs, instruction manuals, and labor to put everything together will double production. In other words, this process can be viewed as production with a fixed cost and a constant marginal cost.

Figure 4.2 plots a production function $y = f(x) = 100 * (x - F)$ that exhibits a fixed cost F and a constant marginal cost of production. Think of y as copies of the next generation of word-processing software with voice recognition (let's call it "WordTalk"), and think of x as the amount of labor input required to produce WordTalk. In this example, F units of labor are required to produce the first copy of WordTalk.[6] Thus, F is the research cost, which is likely to be a very large number. If x is measured as hours of labor input, we might assume that $F = 10{,}000$: it takes 10,000 hours to produce the first copy of WordTalk. After the first copy is created, additional copies can be produced very cheaply. In our example, one hour of labor input can produce 100 copies of the software.

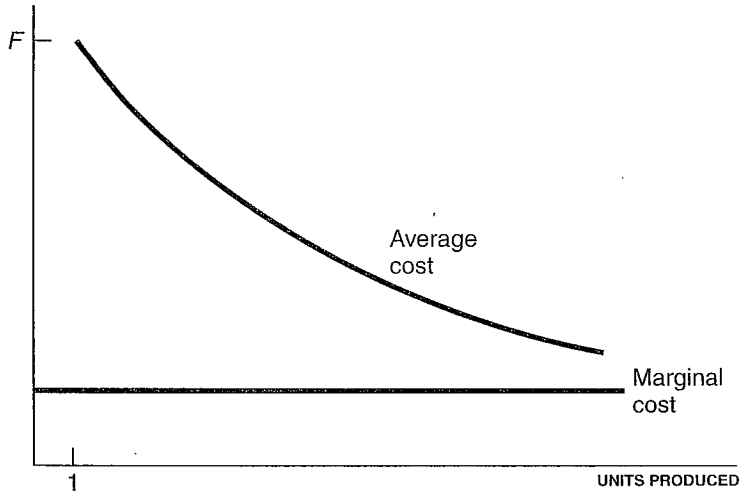
[6]The careful reader will notice that this statement is only approximately right. Actually, $F + 1/100$ units of labor are required to produce the first copy.

Recall that a production function exhibits increasing returns to scale if $f(ax) > af(x)$ where a is some number greater than one — for example, doubling the inputs more than doubles output. Clearly, this is the case for the production function in Figure 4.2. F units of input are required before any output can be produced; $2F$ units of input will produce $100 * F$ units of output. The increasing returns can also be seen in that labor productivity, y/x, is rising with the scale of production.

A common question about software pricing (and the pricing of lots of other goods including CDs, books, and pharmaceuticals) is "If the marginal cost of production is very small, why is it that the product costs so much? Doesn't this imply an inefficiency in the market?" The answer is that yes, there is an inefficiency — remember from your first microeconomics class that efficiency requires that price be equal to marginal cost. However, the inefficiency is in many ways a necessary one.

To explain why, Figure 4.3 shows that the presence of a fixed cost, or more generally the presence of increasing returns, implies that setting price equal to marginal cost will result in negative profits. This figure

FIGURE 4.3 FIXED COSTS AND INCREASING RETURNS



shows the costs of production as a function of the number of units produced. The marginal cost of production is constant — e.g., it costs $10 to produce each additional unit of software. But the average cost is declining. The first unit costs F to produce because of the fixed cost of the idea, which is also the average cost of the first unit. At higher levels of production, this fixed cost is spread over more and more units so that the average cost declines with scale.

Now consider what happens if this firm sets price equal to marginal cost. *With increasing returns to scale, average cost is always greater than marginal cost and therefore marginal cost pricing results in negative profits.* In other words, no firm would enter this market and pay the fixed cost F to develop the computer software if it could not set the price above the marginal cost of producing additional units. In practice, of course, this is exactly what we see: software sells for tens or hundreds of dollars, when the marginal cost of production is presumably only five or ten dollars. Firms will enter only if they can charge a price higher than marginal cost that allows them to recoup the fixed cost of creating the good in the first place. The production of new goods, or new ideas, requires the possibility of earning profits and therefore necessitates a move away from perfect competition.

4.3 INTELLECTUAL PROPERTY RIGHTS AND THE INDUSTRIAL REVOLUTION

In this chapter, we've explained several key features of the economics of ideas. Central among these features is that the economics of ideas involves potentially large one-time costs to create inventions. Think of the cost of creating the first copy of Windows XP or the first jet engine. Inventors will not incur these one-time costs unless they have some expectation of being able to capture some of the gains to society, in the form of profit, after they create the invention. Patents and copyrights are legal mechanisms that grant inventors monopoly power for a time in order to allow them to reap a return from their inventions. They are attempts to use the legal system to influence the degree of excludability of ideas. Without the patent or copyright, it may be quite easy for someone to "reverse engineer" an invention and the competition from this imitation might eliminate the incentive for the inventor to create the idea in the first place. According to some economic historians such as 1993 Nobel laureate Douglass C. North, this reasoning is quite important in understanding the broad history of economic growth, as we will now explain.

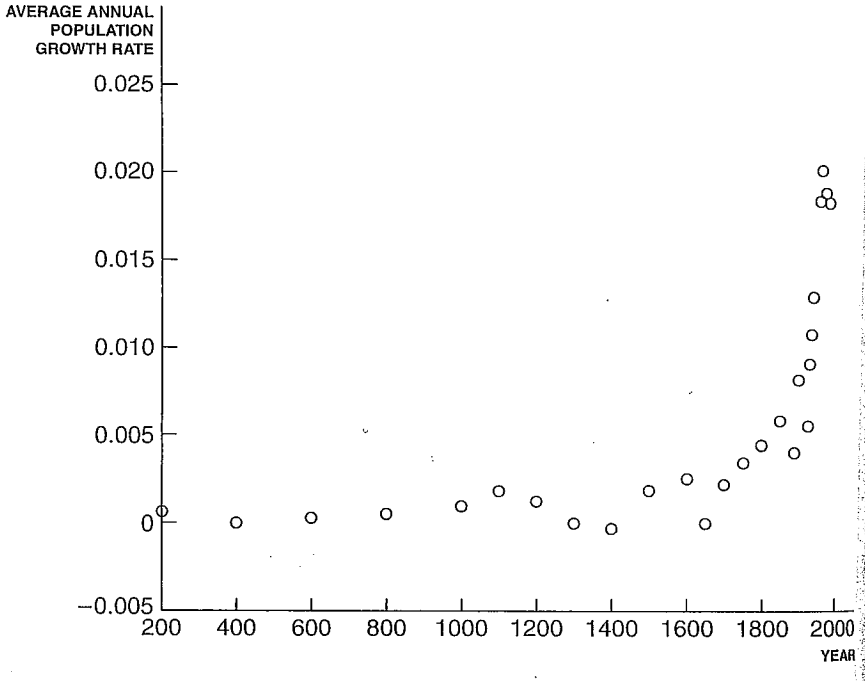
One of the important facts about world economic growth is that it is a very recent phenomenon. Prior to the Industrial Revolution in Britain, the beginning of which historians date to the 1760s, sustained, rapid growth in per capita income was virtually unknown in the world. The general problem with illustrating this point is that we do not have good data on GDP going back much before 1700 or 1800. However, we can exploit the arguments of Thomas Malthus and use population growth to proxy for income growth.[7] That is, for large spans of history, we conjecture that population and income are closely related. For example, the discovery of a new technique in agriculture initially leads to a temporary increase in income, a reduction in mortality, and therefore to an increase in the rate of population growth as more people can be supported by the available land. Gradually, however, diminishing returns to agriculture lead income to fall back to its original (subsistence) level, albeit with a larger population. It is only when sustained increases in per capita income occur that high, sustained rates of population growth are possible.

With this in mind, consider Figure 4.4, which plots average annual rates of world population growth for the last two thousand years. For most of the history of the world, population growth was extremely slow. Indeed, Michael Kremer (1993) reports that the average population growth rate from 1 million B.C. to 1 A.D. was 0.0007 percent per year.[8] From 1 A.D. to 1700, the average annual rate of population growth was still only 0.075 percent per year. During the eighteenth century, population growth rates accelerated, and in the last forty years, world population has grown at an average annual rate of nearly 2 percent per year.

[7] Kremer (1993) provides a detailed application of this technique.

[8] This example illustrates the remarkable power of compounding: even at this near-zero growth rate, world population increased more than a thousandfold over this million-year period.

FIGURE 4.4 **WORLD POPULATION GROWTH, 1 A.D. TO 1990**



SOURCE: Author's calculations and Kremer (1993).

To help to place these numbers in perspective, suppose we were to map out world history on a football field. Let the goal line on one end of the field stand for 1 million B.C., which is a conservative estimate of when humans first became distinguishable from other primates. Let the other goal line correspond to 2000 A.D. Humans were essentially hunters and gatherers for the overwhelming majority of history, until the development of agriculture approximately ten thousand years ago. On our football field, hunting and gathering occupies the first 99 yards of the 100-yard field; systematic agriculture begins on the one-yard line. The year 1 A.D. is only 7 inches from the final goal line, and the Industrial Revolution begins less than one inch from the goal line. In the history of humankind, the era of modern economic growth is the width of a golf ball perched at the end of a football field.

Clearly, sustained economic growth is a very recent phenomenon, and this raises one of the fundamental questions of economic history. How did sustained growth get started in the first place? The thesis of North and a number of other economic historians is that the development of intellectual property rights, a cumulative process that occurred over centuries, is responsible for modern economic growth. It is not until individuals are encouraged by the credible promise of large returns via the marketplace that sustained innovation occurs. To quote a concise statement of this thesis,

> What determines the rate of development of new technology and of pure scientific knowledge? In the case of technological change, the social rate of return from developing new techniques had probably always been high; but we would expect that until the means to raise the private rate of return on developing new techniques was devised, there would be slow progress in producing new techniques.... [T]hroughout man's past he has continually developed new techniques, but the pace has been slow and intermittent. The primary reason has been that the incentives for developing new techniques have occurred only sporadically. Typically, innovations could be copied at no cost by others and without any reward to the inventor or innovator. The failure to develop systematic property rights in innovation up until fairly modern times was a major source of the slow pace of technological change (North 1981, p. 164).

A fascinating and illustrative example of this thesis is provided by the history of navigation. Perhaps the foremost obstacle to the development of ocean shipping, international trade, and world exploration was the problem of determining a ship's location at sea. Latitude was easily discerned by the angle of the North Star above the horizon. However, determining a ship's longitude at sea — its location in the east–west dimension — was a tremendously important problem that remained unsolved until recently. When Columbus landed in the Americas, he thought he had discovered a new route to India because he had no idea of his longitude.

Several astronomical observatories built in western Europe during the seventeenth and eighteenth centuries were sponsored by governments for the express purpose of solving the problem of longitude. The rulers of Spain, Holland, and Britain offered large monetary prizes for the solution. Finally, the problem was solved in the mid-1700s, on the

eve of the Industrial Revolution, by a poorly educated but eminently skilled clockmaker in England named John Harrison. Harrison spent his lifetime building and perfecting a mechanical clock, the chronometer, whose accuracy could be maintained despite turbulence and frequent changes in weather over the course of an ocean voyage that might last for months. This chronometer, rather than any astronomical observation, provided the first practical solution to the determination of longitude.

How does a chronometer solve the problem? Imagine taking two wristwatches with you on a cruise from London to New York. Maintain London (Greenwich!) time on one watch, and set the other watch to noon every day when the sun is directly overhead. The difference in times between the two watches reveals one's longitude relative to the prime meridian.[9]

The lesson of this story for the economist is less in the details of how a chronometer solved the problem of longitude and more in the details of what financial incentives led to the solution. From this standpoint the astounding fact is that there was no *market* mechanism generating the enormous investments required to find a solution. It is not that Harrison or anyone else would become rich from selling the solution to the navies and merchants of western Europe, despite the fact that the benefits to the world from the solution were enormous. Instead, the main financial incentive seems to have been the prizes offered by the governments. Although the Statute of Monopolies in 1624 established a patent law in Britain and the institutions to secure property rights were well on their way in the late eighteenth century, they were still not sufficiently developed to provide the financial incentives for private investment in solving the problem of longitude.[10]

Sustained and rapid economic growth first made its appearance on the world stage during the eighteenth and nineteenth centuries, after literally millions of years of relative stagnation. Exactly why this change occurred remains one of the great mysteries of economics and history. It is tempting to conclude that one of the causes was the establishment of long-lasting institutions that allowed entrepreneurs to capture as

[9]Sobel (1995) discusses the history of longitude in much more detail.
[10]See North and Thomas (1973).

private return some of the enormous social returns their innovations create.[11]

4.4 DATA ON IDEAS

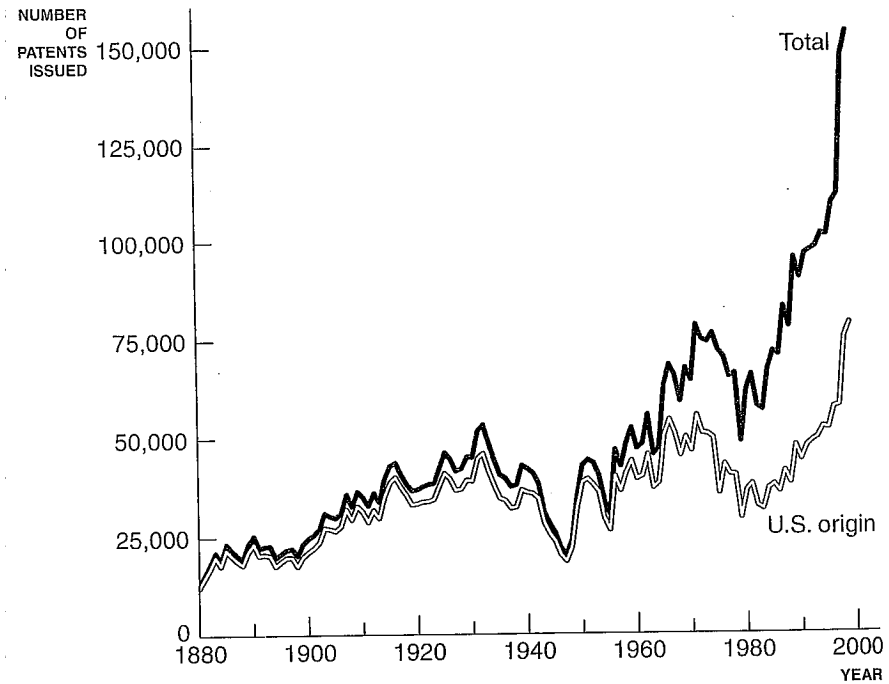
What data do we have on ideas? At some fundamental level it is difficult to measure both the inputs to the production function for ideas and the output of that production function, the ideas themselves. At the same time, data that correspond roughly to both the inputs and the output do exist. For example, R&D is presumably a very important input into the production function for ideas. To the extent that the most important or valuable ideas are patented, patent counts may provide a simple measure of the number of ideas produced. Of course, both of these measures have their problems. Many ideas are neither patented nor produced using resources that are officially labelled as R&D. The Wal-Mart operation manual and multiplex movie theaters are good examples. In addition, a simple count of the number of patents granted in any particular year does not convey the economic value of the patents. Among the thousands of patents awarded every year, only one may be for the transistor or the laser.

Nevertheless, let us examine the patent and R&D data, keeping these caveats in mind. A patent is a legal document that describes an invention and entitles the patent owner to a monopoly over the invention for some period of time, typically 17 to 20 years. Figure 4.5 plots the number of patents awarded in every year from 1880 until 1999. The first feature apparent from the graph is the rise in the number of patents awarded. In 1880, approximately 13,000 patents were issued; in 1999, more than 150,000 patents were issued. Presumably, the number of ideas used in the U.S. economy increased substantially over the century.

This large increase masks several important features of the data, however. First, nearly half of all patents granted in 1999 were of foreign

[11]The confluence of events in the late eighteenth century is remarkable and suggestive of a broader set of causes. In addition to the beginning of the Industrial Revolution, we have the drafting of the Declaration of Independence, the U.S. Constitution and the Bill of Rights, the French Declaration of the Rights of Man and of the Citizen, and the publication of Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations.*

FIGURE 4.5 PATENTS ISSUED IN THE UNITED STATES, 1880–1999

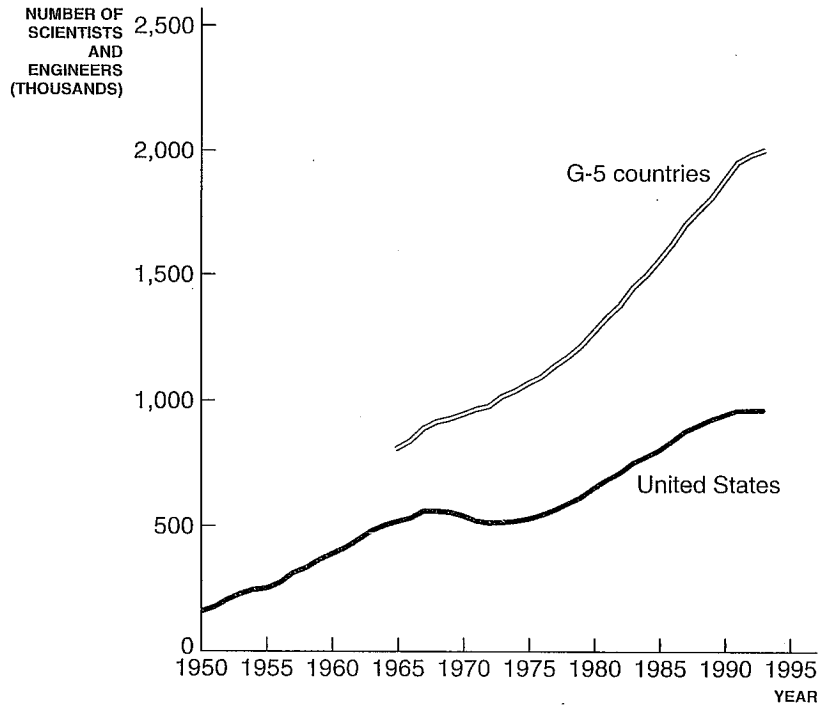


SOURCE: U.S. Patent and Trademark Office (2001).

origin. Second, nearly all of the increase in patents over the last century reflects an increase in foreign patents, at least until the 1990s; the number of patents awarded in the United States to U.S. residents was around 40,000 in 1915, 1950, and 1988. Does this mean that the number of new ideas generated within the United States has been relatively constant from 1915 to the present? Probably not. It is possible that the value of patents has increased or that fewer new ideas are patented. The formula for Coca-Cola, for example, is a quietly kept trade secret that has never been patented.

What about the inputs into the production of ideas? Figure 4.6 plots the number of scientists and engineers engaged in R&D from 1950 to 1993. During this forty-year period, resources devoted to R&D increased

FIGURE 4.6 SCIENTISTS AND ENGINEERS ENGAGED IN R&D, 1950–93



SOURCE: Jones (2001). The G-5 countries are France, Germany (West Germany until 1990), Japan, the United Kingdom, and the United States.

dramatically in the United States, from less than 200,000 scientists and engineers in 1950 to nearly 1 million in 1993. A similar rise can be seen for the five most highly developed countries as a whole.

Not only has the *level* of resources devoted to R&D increased, but the *share* of resources devoted to R&D has also increased. The number of U.S. scientists and engineers engaged in R&D increased from about 0.25 percent of the labor force in 1950 to around 0.75 percent in 1993. The numbers are similarly striking for Japan, France, Germany, and the United Kingdom. For example, the share in Japan rose from 0.2 percent in 1965 to nearly 0.8 percent in 1990.

SUMMARY

One of the main contributions of new growth theory has been to emphasize that ideas are very different from other economic goods. Ideas are nonrivalrous: once an idea is invented, it can be used by one person or by one thousand people, at no additional cost. This distinguishing feature of ideas implies that the size of the economy — its scale — plays an important role in the economics of ideas. In particular, the nonrivalry of ideas implies that production will be characterized by increasing returns to scale. In turn, the presence of increasing returns suggests that we must move away from models of perfect competition. The only reason an inventor is willing to undertake the large one-time costs of creating a new idea is because the inventor expects to be able to charge a price greater than marginal cost and earn profits.

New ideas often create benefits that the inventor is unable to capture. This is what is meant when we say that ideas are only partially excludable. The incentive to create new ideas depends on the profits that an inventor can expect to earn (the private benefit), not on the entire social benefit generated by the idea. Whether or not an idea gets created depends on the magnitude of the private benefit relative to the one-time invention costs. It is easy to see, then, how ideas that are socially very valuable may fail to be invented if private benefits and social benefits are too far apart. Patents and copyrights are legal mechanisms that attempt to bring the private benefits of invention closer in line with the social benefits. The development of such institutions — and of property rights more generally — may have played a critical role in sparking the Industrial Revolution and the sustained economic growth that has followed.

EXERCISES

1. *Classifying goods.* Place the following goods on a chart like that in Figure 4.1 — i.e., classify them as rivalrous or nonrivalrous and by the extent to which they are excludable: a chicken, the trade secret for Coca-Cola, music from a compact disc, tropical rainforests, clean air, and a lighthouse that guides ships around a rocky coast.

2. *Provision of goods.* Explain the role of the market and the government in providing each of the goods in the previous question.

3. *Pricing with increasing returns to scale.* Consider the following production function (similar to that used earlier for WordTalk):

$$Y = 100 * (L - F)$$

where Y is output, L is labor input, and F is a fixed amount of labor that is required before the first unit of output can be produced (like a research cost). We assume that $Y = 0$ if $L < F$. Each unit of labor L costs the wage w to hire.

(a) How much does it cost (in terms of wages) to produce 5 units of output?

(b) More generally, how much does it cost to produce *any* arbitrary amount of output, Y? That is, find the cost function $C(Y)$ that tells the minimum cost required to produce Y units of output.

(c) Show that the marginal cost dC/dY is constant (after the first unit is produced).

(d) Show that the average cost C/Y is declining.

(e) Show that if the firm charges a price P equal to marginal cost, its profits, defined as $\pi = PY - C(Y)$, will be negative regardless of the level of Y.

5 THE ENGINE OF GROWTH

> As for the Arts of Delight and Ornament, they are best promoted by the greatest number of emulators. And it is more likely that one ingenious curious man may rather be found among 4 million than among 400 persons....
>
> —WILLIAM PETTY, (cited in Simon (1981), p. 158).

The neoclassical growth model highlights technological progress as the engine of economic growth, and the previous chapter discussed in broad terms the economics of ideas and technology. In this chapter, we incorporate the insights from the previous chapters to develop an explicit theory of technological progress. The model we develop allows us to explore the engine of economic growth, thus addressing the second main question posed at the beginning of this book. We seek an understanding of why the advanced economies of the world, such as the United States, have grown at something like 2 percent per year for the last century. Where does the technological progress that underlies this growth come from? Why is the growth rate 2 percent per year instead of 1 percent or 10 percent? Can we expect this growth to continue, or is there some limit to economic growth?

Much of the work by economists to address these questions has been labeled *endogenous growth theory* or *new growth theory*. Instead of assuming that growth occurs because of automatic and unmodeled (exogenous) improvements in technology, the theory focuses on understanding the economic forces underlying technological progress. An important contribution of this work is the recognition that technological progress occurs as profit-maximizing firms or inventors seek out newer and better mousetraps. Adam Smith wrote that "it is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest" (Smith 1776 [1981], pp. 26–7). Similarly, it is the possibility of earning a profit that drives firms to develop a computer that can fit in your hand, a soft drink with only a single calorie, or a way to record TV programs and movies to be replayed at your convenience. In this way, improvements in technology, and the process of economic growth itself, are understood as an endogenous outcome of the economy.

The specific theory we will develop in this chapter was constructed by Paul Romer in a series of papers, including a 1990 paper titled "Endogeneous Technological Change."[1]

5.1 THE BASIC ELEMENTS OF THE MODEL

The Romer model endogenizes technological progress by introducing the search for new ideas by researchers interested in profiting from their inventions. The market structure and economic incentives that are at the heart of this process will be examined in detail in Section 5.2. First, though, we will outline the basic elements of the model and their implications for economic growth.

The model is designed to explain why and how the advanced countries of the world exhibit sustained growth. In contrast to the neoclassical models in earlier chapters, which could be applied to different countries, the model in this chapter describes the advanced countries of the world as a whole. Technological progress is driven by research and development (R&D) in the advanced world. In the next chapter we

[1]The version of the Romer model that we will present in this chapter is based on Jones (1995a). There is one key difference between the two models, which will be discussed at the appropriate time. Other notable contributions to the literature on R&D-based growth models include Grossman and Helpman (1991) and Aghion and Howitt (1992). These models are sometimes called Schumpeterian growth models, because they were anticipated by the work of Joseph Schumpeter in the late 1930s and early 1940s.

will explore the important process of technology transfer and why different economies have different levels of technology. For the moment, we will concern ourselves with how the world technological frontier is continually pushed outward.

As was the case with the Solow model, there are two main elements in the Romer model of endogenous technological change: an equation describing the production function and a set of equations describing how the inputs for the production function evolve over time. The main equations will be similar to the equations for the Solow model, with one important difference.

The aggregate production function in the Romer model describes how the capital stock, K, and labor, L_Y, combine to produce output, Y, using the stock of ideas, A:

$$Y = K^{\alpha}(AL_Y)^{1-\alpha}, \tag{5.1}$$

where α is a parameter between 0 and 1. For the moment, we take this production function as given; in Section 5.2, we will discuss in detail the market structure and the microfoundations of the economy that underlie this aggregate production function.

For a given level of technology, A, the production function in equation (5.1) exhibits constant returns to scale in K and L_Y. However, when we recognize that ideas (A) are also an input into production, then there are increasing returns. For example, once Steve Jobs and Steve Wozniak invented the plans for assembling personal computers, those plans (the "idea") did not need to be invented again. To double the production of personal computers, Jobs and Wozniak needed only to double the number of integrated circuits, semiconductors, etc., and find a larger garage. That is, the production function exhibits constant returns to scale with respect to the capital and labor inputs, and therefore must exhibit increasing returns with respect to all three inputs: if you double capital, labor, *and* the stock of ideas, then you will more than double output. As discussed in Chapter 4, the presence of increasing returns to scale results fundamentally from the nonrivalrous nature of ideas.

The accumulation equations for capital and labor are identical to those for the Solow model. Capital accumulates as people in the economy forego consumption at some given rate, s_K, and depreciates at the exogenous rate d:

$$\dot{K} = s_K Y - dK.$$

Labor, which is equivalent to the population, grows exponentially at some constant and exogenous rate n:

$$\frac{\dot{L}}{L} = n.$$

The key equation that is new relative to the neoclassical model is the equation describing technological progress. In the neoclassical model, the productivity term A grows exogenously at a constant rate. In the Romer model, growth in A is endogenized. How is this accomplished? The answer is with a production function for new ideas: just as more automobile workers can produce more cars, we assume that more researchers can produce more new ideas.

According to the Romer model, $A(t)$ is the stock of knowledge or the number of ideas that have been invented over the course of history up until time t. Then, $\dot{A}$ is the number of new ideas produced at any given point in time. In the simplest version of the model, $\dot{A}$ is equal to the number of people attempting to discover new ideas, L_A, multiplied by the rate at which they discover new ideas, $\bar{\delta}$:

$$\dot{A} = \bar{\delta} L_A. \tag{5.2}$$

The rate at which researchers discover new ideas might simply be a constant. On the other hand, one could imagine that it depends on the stock of ideas that have already been invented. For example, perhaps the invention of ideas in the past raises the productivity of researchers in the present. In this case, $\bar{\delta}$ would be an increasing function of A. The discovery of calculus, the invention of the laser, and the development of integrated circuits are examples of ideas that have increased the productivity of later research. On the other hand, perhaps the most obvious ideas are discovered first and subsequent ideas are increasingly difficult to discover. In this case, $\bar{\delta}$ would be a decreasing function of A.

This reasoning suggests modeling the rate at which new ideas are produced as

$$\bar{\delta} = \delta A^{\phi}, \tag{5.3}$$

where δ and ϕ are constants. In this equation, $\phi > 0$ indicates that the productivity of research increases with the stock of ideas that have

already been discovered; $\phi < 0$ corresponds to the "fishing out" case in which the fish become harder to catch over time. Finally, $\phi = 0$ indicates that the tendency for the most obvious ideas to be discovered first exactly offsets the fact that old ideas may facilitate the discovery of new ideas—i.e., the productivity of research is independent of the stock of knowledge.

It is also possible that the average productivity of research depends on the number of people searching for new ideas at any point in time. For example, perhaps duplication of effort is more likely when there are more persons engaged in research. One way of modeling this possibility is to suppose that it is really L_A^{λ}, where λ is some parameter between 0 and 1, rather than L_A that enters the production function for new ideas. This, together with equations (5.3) and (5.2), suggests focusing on the following general production function for ideas:

$$\dot{A} = \delta L_A^{\lambda} A^{\phi}. \tag{5.4}$$

For reasons that will become clear, we will assume that $\phi < 1$.

Equations (5.2) and (5.4) illustrate a very important aspect of modeling economic growth.[2] Individual researchers, being small relative to the economy as a whole, take $\bar{\delta}$ as given and see constant returns to research. As in equation (5.2), an individual engaged in research creates $\bar{\delta}$ new ideas. In the economy as a whole, however, the production function for ideas may not be characterized by constant returns to scale. While $\bar{\delta}$ will change by only a minuscule amount in response to the actions of a single researcher, it clearly varies with aggregate research effort.[3] For example, $\lambda < 1$ may reflect an externality associated with duplication: some of the ideas created by an individual researcher may not be new to the economy as a whole. This is analogous to congestion on a highway. Each driver ignores the fact that his or her presence makes it slightly harder for other drivers to get where they are going. The effect of any single driver is negligible, but summed across all drivers, the effects can be important.

[2]This modeling technique will be explored again in Chapter 8 in the context of "AK" models of growth.

[3]Notice that the exact expression for $\bar{\delta}$, incorporating both duplication and knowledge spillovers, is $\bar{\delta} = \delta L_A^{\lambda-1} A^{\phi}$.

Similarly, the presence of A^{ϕ} is treated as external to the individual agent. Consider the case of $\phi > 0$, reflecting a positive knowledge spillover in research. The gains to society from the theory of gravitation far outweighed the benefit that Isaac Newton was able to capture. Much of the knowledge he created "spilled over" to future researchers. Of course, Newton himself also benefited from the knowledge created by previous scientists such as Kepler, as he recognized in the famous statement, "If I have seen farther than others, it is because I was standing on the shoulders of giants." With this in mind, we might refer to the externality associated with ϕ as the "standing on shoulders" effect, and by extension, the externality associated with λ as the "stepping on toes" effect. :-)

Next, we need to discuss how resources are allocated in this economy. There are two key allocations. First, we assume (as before) that a constant fraction of output is invested in capital. Second, we have to decide how much labor works to produce output and how much works to produce ideas, recognizing that these two activities employ all of the labor in the economy:

$$L_Y + L_A = L.$$

In a more sophisticated model (and indeed, in Romer's original paper), the allocation of labor is determined by utility maximization and markets. However, it is again convenient to make the Solow-style assumption that the allocation of labor is constant; this assumption will be relaxed in Section 5.2. We assume that a constant fraction, $L_A/L = s_R$, of the labor force engages in R&D to produce new ideas, and the remaining fraction, $1 - s_R$, produces output.

Finally, the economy has some initial endowments when it begins. We assume the economy starts out with K_0 units of capital, L_0 units of labor, and A_0 ideas. This completes our setup of the model and we are ready to begin solving for some key endogenous variables, beginning with the long-run growth rate of this economy.

5.1.1 GROWTH IN THE ROMER MODEL

What is the growth rate in this model along a balanced growth path? Provided a constant fraction of the population is employed producing ideas (which we will show to be the case below), the model follows

the neoclassical model in predicting that all per capita growth is due to technological progress. Letting lower-case letters denote per capita variables, and letting g_x denote the growth rate of some variable x along the balanced growth path, it is easy to show that

$$g_y = g_k = g_A.$$

That is, per capita output, the capital-labor ratio, and the stock of ideas must all grow at the same rate along a balanced growth path.[4] If there is no technological progress in the model, then there is no growth.

Therefore, the important question is "What is the rate of technological progress along a balanced growth path?" The answer to this question is found by rewriting the production function for ideas, equation (5.4). Dividing both sides of this equation by A yields

$$\frac{\dot{A}}{A} = \delta \frac{L_A^{\lambda}}{A^{1-\phi}}. \qquad (5.5)$$

Along a balanced growth path, $\dot{A}/A \equiv g_A$ is constant. But this growth rate will be constant if and only if the numerator and the denominator of the right-hand side of equation (5.5) grow at the same rate. Taking logs and derivatives of both sides of this equation,

$$0 = \lambda \frac{\dot{L}_A}{L_A} - (1-\phi)\frac{\dot{A}}{A}. \qquad (5.6)$$

Along a balanced growth path, the growth rate of the number of researchers must be equal to the growth rate of the population — if it were higher, the number of researchers would eventually exceed the population, which is impossible. That is, $\dot{L}_A/L_A = n$. Substituting this into equation (5.6) yields

$$g_A = \frac{\lambda n}{1-\phi}. \qquad (5.7)$$

[4]To see this, follow the arguments we made in deriving equation (2.10) in Chapter 2. Intuitively, the capital-output ratio must be constant along a balanced growth path. Recognizing this fact, the production function implies that y and k must grow at the same rate as A.

Thus the long-run growth rate of this economy is determined by the parameters of the production function for ideas and the rate of growth of researchers, which is ultimately given by the population growth rate.

Several features of this equation deserve comment. First, what is the intuition for the equation? The intuition is most easily seen by considering the special case in which $\lambda = 1$ and $\phi = 0$ so that the productivity of researchers is the constant δ. In this case, there is no duplication problem in research and the productivity of a researcher today is independent of the stock of ideas that have been discovered in the past. The production function for ideas looks like

$$\dot{A} = \delta L_A.$$

Now suppose that the number of people engaged in the search for ideas is constant. Because δ is also constant, this economy generates a constant number of new ideas, δL_A, each period. To be more concrete, let's suppose $\delta L_A = 100$. The economy begins with some stock of ideas, A_0, generated by previous discoveries. Initially, the 100 new ideas per period may be a large fraction of the existing stock, A_0. Over time, though, the stock grows, and the 100 new ideas becomes a smaller and smaller fraction of the existing stock. Therefore, the *growth rate* of the stock of ideas falls over time, eventually approaching zero. Notice, however, that technological progress never ceases. The economy is always creating 100 new ideas. It is simply that these 100 new ideas shrink in comparison with the accumulated stock of ideas.

In order to generate exponential growth, the number of new ideas must be expanding over time. This occurs if the number of researchers is increasing — for example, because of world population growth. More researchers mean more ideas, sustaining growth in the model. In this case, the growth in ideas is clearly related to the growth in population, which explains the presence of population growth in equation (5.7). Phelps (1968) clarifies the intuition for this basic result with an enlightening example:

> One can hardly imagine, I think, how poor we would be today were it not for the rapid population growth of the past to which we owe the enormous number of technological advances enjoyed today.... If I could re-do the history of the world, halving population size each year from the beginning of time on some random basis, I would not do it for fear of losing Mozart in the process (pp. 511–512).

It is interesting to compare this result to the effect of population growth in the neoclassical growth model. There, for example, a higher population growth rate reduces the level of income along a balanced growth path. More people means that more capital is needed to keep K/L constant, but capital runs into diminishing returns. Here, an important additional effect exists. People are the key input to the creative process. A larger population generates more ideas, and because ideas are nonrivalrous, everyone in the economy benefits.

What evidence can be presented to support the contention that the per capita growth rate of the world economy depends on population growth? First, notice that this particular implication of the model is very difficult to test. We have already indicated that this model of the engine of growth is meant to describe the advanced countries of the world taken as a whole. Thus, we cannot use evidence on population growth *across* countries to test the model. In fact, we have already presented one of the most compelling pieces of evidence in Chapter 4. Recall the plot in Figure 4.4 of world population growth rates over the last 2,000 years. Sustained and rapid population growth is a rather recent phenomenon, just as is sustained and rapid growth in per capita output. Increases in the rate of population growth from the very low rate observed over most of history occurred at roughly the same time as the Industrial Revolution.

The result that the growth rate of the economy is tied to the growth rate of the population implies another seemingly strong result: if the population (or at least the number of researchers) stops growing, long-run growth ceases. What do we make of this prediction? Rephrasing the question slightly, if research effort in the world were constant over time, would economic growth eventually grind to a halt? This model suggests that it would. A constant research effort cannot continue the proportional increases in the stock of ideas needed to generate long-run growth.

Actually, there is one special case in which a constant research effort can sustain long-run growth, and this brings us to our second main comment about the model. The production function for ideas considered in the original Romer (1990) paper assumes that $\lambda = 1$ and $\phi = 1$. That is,

$$\dot{A} = \delta L_A A.$$

Rewriting the equation slightly, we can see that this version of the Romer model *will* generate sustained growth in the presence of a constant research effort:

$$\frac{\dot{A}}{A} = \delta L_A. \tag{5.8}$$

In this case, Romer assumes that the productivity of research is proportional to the existing stock of ideas: $\delta = \delta A$. With this assumption, the productivity of researchers grows over time, even if the number of researchers is constant.

The advantage of this specification, however, is also its drawback. World research effort has increased enormously over the last forty years and even over the last century (see Figure 4.6 in Chapter 4 for a reminder of this fact). Since L_A is growing rapidly over time, the original Romer formulation in equation (5.8) predicts that the growth rate of the advanced economies should also have risen rapidly over the last forty years or the last century. We know this is far from the truth. The average growth rate of the U.S. economy, for example, has been very close to 1.8 percent per year for the last hundred years. This easily rejected prediction of the original Romer formulation is avoided by requiring that ϕ is less than one, which returns us to the results associated with equation (5.7).[5]

Notice that nothing in this reasoning rules out increasing returns in research or positive knowledge spillovers. The knowledge spillover parameter, ϕ, may be positive and quite large. What the reasoning points out is that the somewhat arbitrary case of $\phi = 1$ is strongly rejected by empirical observation.[6]

Our last comment about the growth implications of this model of technology is that the results are similar to the neoclassical model in one important way. In the neoclassical model, changes in government policy and changes in the investment rate have no long-run effect on economic growth. This result was not surprising once we recognized that all growth in the neoclassical model was due to exogenous technological progress. In this model with endogenous technological progress, however, we have the same result. The long-run growth rate is invari-

[5]This point is made in Jones (1995a).

[6]The same evidence also rules out values of $\phi > 1$. Such values would generate accelerating growth rates even with a constant population!

It is interesting to compare this result to the effect of population growth in the neoclassical growth model. There, for example, a higher population growth rate reduces the level of income along a balanced growth path. More people means that more capital is needed to keep K/L constant, but capital runs into diminishing returns. Here, an important additional effect exists. People are the key input to the creative process. A larger population generates more ideas, and because ideas are nonrivalrous, everyone in the economy benefits.

What evidence can be presented to support the contention that the per capita growth rate of the world economy depends on population growth? First, notice that this particular implication of the model is very difficult to test. We have already indicated that this model of the engine of growth is meant to describe the advanced countries of the world taken as a whole. Thus, we cannot use evidence on population growth *across* countries to test the model. In fact, we have already presented one of the most compelling pieces of evidence in Chapter 4. Recall the plot in Figure 4.4 of world population growth rates over the last 2,000 years. Sustained and rapid population growth is a rather recent phenomenon, just as is sustained and rapid growth in per capita output. Increases in the rate of population growth from the very low rate observed over most of history occurred at roughly the same time as the Industrial Revolution.

The result that the growth rate of the economy is tied to the growth rate of the population implies another seemingly strong result: if the population (or at least the number of researchers) stops growing, long-run growth ceases. What do we make of this prediction? Rephrasing the question slightly, if research effort in the world were constant over time, would economic growth eventually grind to a halt? This model suggests that it would. A constant research effort cannot continue the proportional increases in the stock of ideas needed to generate long-run growth.

Actually, there is one special case in which a constant research effort can sustain long-run growth, and this brings us to our second main comment about the model. The production function for ideas considered in the original Romer (1990) paper assumes that $\lambda = 1$ and $\phi = 1$. That is,

$$\dot{A} = \delta L_A A. \tag{5.7}$$

Rewriting the equation slightly, we can see that this version of the Romer model *will* generate sustained growth in the presence of a constant research effort:

$$\frac{\dot{A}}{A} = \delta L_A. \tag{5.8}$$

In this case, Romer assumes that the productivity of research is proportional to the existing stock of ideas: $\delta = \delta A$. With this assumption, the productivity of researchers grows over time, even if the number of researchers is constant.

The advantage of this specification, however, is also its drawback. World research effort has increased enormously over the last forty years and even over the last century (see Figure 4.6 in Chapter 4 for a reminder of this fact). Since L_A is growing rapidly over time, the original Romer formulation in equation (5.8) predicts that the growth rate of the advanced economies should also have risen rapidly over the last forty years or the last century. We know this is far from the truth. The average growth rate of the U.S. economy, for example, has been very close to 1.8 percent per year for the last hundred years. This easily rejected prediction of the original Romer formulation is avoided by requiring that ϕ is less than one, which returns us to the results associated with equation (5.7).[5]

Notice that nothing in this reasoning rules out increasing returns in research or positive knowledge spillovers. The knowledge spillover parameter, ϕ, may be positive and quite large. What the reasoning points out is that the somewhat arbitrary case of $\phi = 1$ is strongly rejected by empirical observation.[6]

Our last comment about the growth implications of this model of technology is that the results are similar to the neoclassical model in one important way. In the neoclassical model, changes in government policy and changes in the investment rate have no long-run effect on economic growth. This result was not surprising once we recognized that all growth in the neoclassical model was due to exogenous technological progress. In this model with endogenous technological progress, however, we have the same result. The long-run growth rate is invari-

[5]This point is made in Jones (1995a).

[6]The same evidence also rules out values of $\phi > 1$. Such values would generate accelerating growth rates even with a constant population!

ant to changes in the investment rate, and even to changes in the share of the population that is employed in research. This is seen by noting that none of the parameters in equation (5.7) is affected when, say, the investment rate or the R&D share of labor is changed. Instead, these policies affect the growth rate along a transition path to the new steady state altering the *level* of income. That is, even after we endogenize technology in this model, the long-run growth rate cannot be manipulated by policy makers using conventional policies such as subsidies to R&D.

5.1.2 GROWTH EFFECTS VERSUS LEVEL EFFECTS

The fact that standard policies cannot affect long-run growth is *not* a feature of the original Romer model, nor of many other idea-based growth models that followed, including Grossman and Helpman (1991) and Aghion and Howitt (1992). Much of the theoretical work in new growth theory has sought to develop models in which policy changes *can* have effects on long-run growth.

The idea-based models in which changes in policy can permanently increase the growth rate of the economy all rely on the assumption that $\phi = 1$, or its equivalent. As shown above, this assumption generates the counterfactual prediction that growth rates should accelerate over time with a growing population. Jones (1995a) generalized these models to the case of $\phi < 1$ to eliminate this defect and showed the somewhat surprising implication that this eliminates the long-run growth effects of policy as well. We will discuss these issues in more detail in Chapter 8.

5.1.3 COMPARATIVE STATICS: A PERMANENT INCREASE IN THE R&D SHARE

What happens to the advanced economies of the world if the share of the population searching for new ideas increases permanently? For example, suppose there is a government subsidy for R&D that increases the fraction of the labor force doing research.

An important feature of the model we have just developed is that many policy changes (or comparative statics) can be analyzed with techniques we have already developed. Why? Notice that technological progress in the model can be analyzed by itself—it doesn't depend

on capital or output, but only on the labor force and the share of the population devoted to research. Once the growth rate of A is constant, the model behaves just like the Solow model with exogenous technological progress. Therefore, our analysis proceeds in two steps. First, we consider what happens to technological progress and to the stock of ideas after the increase in R&D intensity occurs. Second, we analyze the model as we did the Solow model, in steps familiar from Chapter 2. Before we proceed, it is worth noting that the analysis of changes that do not affect technology, such as an increase in the investment rate, is exactly like the analysis of the Solow model.

Now consider what happens if the share of the population engaged in research increases permanently. To simplify things slightly, let's assume that $\lambda = 1$ and $\phi = 0$ again; none of the results are qualitatively affected by this assumption. It is helpful to rewrite equation (5.5) as

$$\frac{\dot{A}}{A} = \delta \frac{s_R L}{A}, \tag{5.9}$$

where s_R is the share of the population engaged in R&D, so that $L_A = s_R L$.

Figure 5.1 shows what happens to technological progress when s_R increases permanently to s_R', assuming the economy begins in steady

FIGURE 5.1 TECHNOLOGICAL PROGRESS: AN INCREASE IN THE R&D SHARE

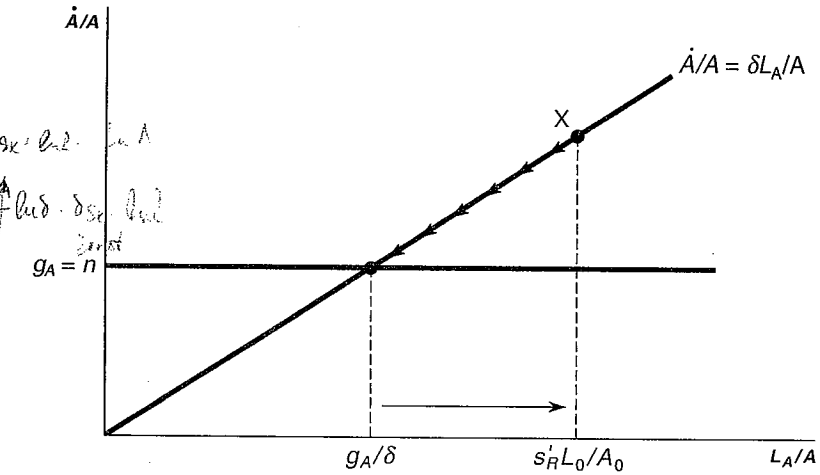
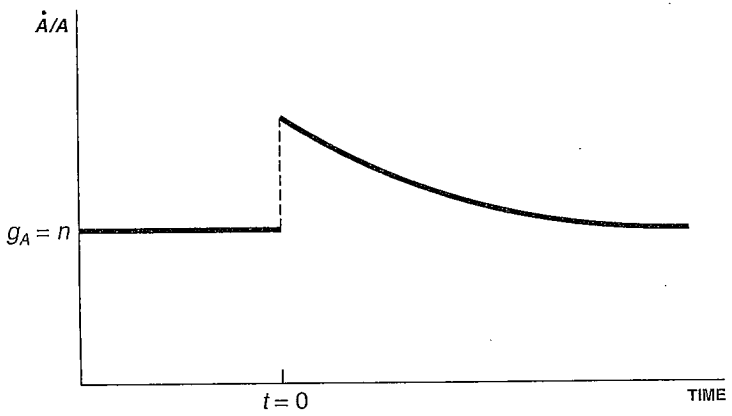


FIGURE 5.2 $\dot{A}/A$ OVER TIME



state. In steady state, the economy grows along a balanced growth path at the rate of technological progress, g_A, which happens to equal the rate of population growth under our simplifying assumptions. The ratio L_A/A is therefore equal to g_A/δ. Suppose the increase in s_R occurs at time $t = 0$. With a population of L_0, the number of researchers increases as s_R increases, so that the ratio L_A/A jumps to a higher level. The additional researchers produce an increased number of new ideas, so the growth rate of technology is also higher at this point. This situation corresponds to the point labeled "X" in the figure. At X, technological progress $\dot{A}/A$ exceeds population growth n, so the ratio L_A/A declines over time, as indicated by the arrows. As this ratio declines, the rate of technological change gradually falls also, until the economy returns to the balanced growth path where $g_A = n$. Therefore, a permanent increase in the share of the population devoted to research raises the rate of technological progress temporarily, but not in the long run. This behavior is depicted in Figure 5.2.

What happens to the level of technology in this economy? Figure 5.3 answers this question. The level of technology is growing along a balanced growth path at rate g_A until time $t = 0$. At this time, the growth rate increases and the level of technology rises faster than before. Over

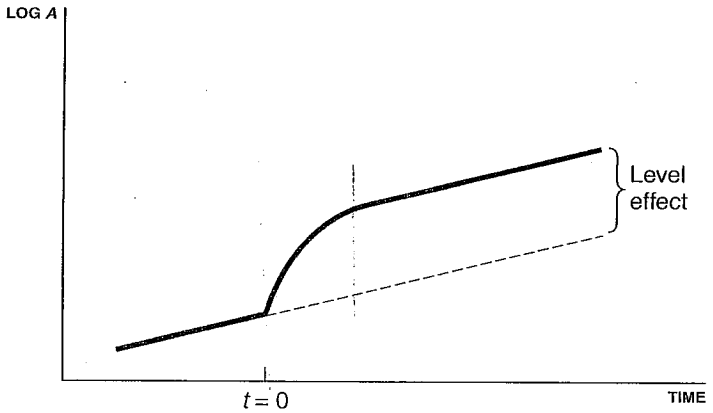
time, however, the growth rate falls until it returns to g_A. The *level* of technology is permanently higher as a result of the permanent increase in R&D. Notice that a permanent increase in s_R in the Romer model generates transition dynamics that are qualitatively similar to the dynamics generated by an increase in the investment rate in the Solow model.

Now that we know what happens to technology over time, we can analyze the remainder of the model in a Solow framework. The long-run growth rate of the model is constant, so much of the algebra that we used in analyzing the Solow model applies. For example, the ratio y/A is constant along a balanced growth path and is given by an equation similar to equation (2.13):

$$\left(\frac{y}{A}\right)^* = \left(\frac{s_K}{n + g_A + d}\right)^{\alpha/(1-\alpha)} (1 - s_R). \tag{5.10}$$

The only difference is the presence of the term $1 - s_R$, which adjusts for the difference between output per worker, L_Y, and output per capita, L. Notice that along a balanced growth path, equation (5.9) can be solved for the level of A in terms of the labor force:

FIGURE 5.3 THE LEVEL OF TECHNOLOGY OVER TIME



Notice that along a balanced growth path, equation (5.9) can be solved for the level of A in terms of the labor force:

$$A = \frac{\delta s_R L}{g_A}.$$

Combining this equation with (5.10), we get

$$y^*(t) = \left(\frac{s_K}{n + g_A + d}\right)^{\alpha/(1-\alpha)} (1 - s_R)\frac{\delta s_R}{g_A} L(t). \qquad (5.11)$$

In this simple version of the model, per capita output is proportional to the population of the (world) economy along a balanced growth path. In other words, the model exhibits a *scale effect* in levels: a larger world economy will be a richer world economy. This scale effect arises fundamentally from the nonrivalrous nature of ideas: a larger economy provides a larger market for an idea, raising the return to research (a demand effect). In addition, a more populous world economy simply has more potential creators of ideas in the first place (a supply effect).

The other terms in equation (5.11) are readily interpreted. The first term is familiar from the original Solow model. Economies that invest more in capital will be richer, for example. Two terms involve the share of labor devoted to research, s_R. The first time s_R appears, it enters negatively to reflect the fact that more researchers mean fewer workers producing output. The second time, it enters positively to reflect the fact that more researchers mean more ideas, which increases the productivity of the economy.

5.2 THE ECONOMICS OF THE MODEL

The first half of this chapter has analyzed the Romer model without discussing the economics underlying the model. A number of economists in the 1960s developed models with similar macroeconomic features.[7] However, the development of the microfoundations of such models had to wait until the 1980s when economists better understood how to model imperfect competition in a general equilibrium setting.[8] In

[7]For example, Uzawa (1965), Phelps (1966), Shell (1967), and Nordhaus (1969).

[8]Key steps in this understanding were accomplished by Spence (1976), Dixit and Stiglitz (1977), and Ethier (1982).

fact, one of the important contributions of Romer (1990) was to explain exactly how to construct an economy of profit-maximizing agents that endogenizes technological progress. The intuition behind this insight was developed in Chapter 4. Developing the mathematics is the subject of the remainder of this section. Because this section is somewhat difficult, some readers may wish to skip to Section 5.3.

The Romer economy consists of three sectors: a final-goods sector, an intermediate-goods sector, and a research sector. The reason for two of the sectors should be clear: some firms must produce output and some firms must produce ideas. The reason for the intermediate-goods sector is related to the presence of increasing returns discussed in Chapter 4. Each of these sectors will be discussed in turn. Briefly, the research sector creates new ideas, which take the form of new varieties of capital goods—new computer chips, fax machines, or printing presses. The research sector sells the exclusive right to produce a specific capital good to an intermediate-goods firm. The intermediate-goods firm, as a monopolist, manufactures the capital good and sells it to the final-goods sector, which produces output.

5.2.1 THE FINAL-GOODS SECTOR

The final-goods sector of the Romer economy is very much like the final-goods sector of the Solow model. It consists of a large number of perfectly competitive firms that combine labor and capital to produce a homogeneous output good, Y. The production function is specified in a slightly different way, though, to reflect the fact that there is more than one capital good in the model:

$$Y = L_Y^{1-\alpha} \sum_{j=1}^{A} x_j^{\alpha}.$$

Output, Y, is produced using labor, L_Y, and a number of different capital goods, x_j, which we will also call "intermediate goods." At any point in time, A measures the number of capital goods that are available to be used in the final-goods sector, and firms in the final-goods sector take this number as given. Inventions or ideas in the model correspond to the creation of new capital goods that can be used by the final-goods sector to produce output.

Notice that this production function can be rewritten as

$$Y = {L_Y}^{1-\alpha}{x_1}^{\alpha} + {L_Y}^{1-\alpha}{x_2}^{\alpha} + \cdots + {L_Y}^{1-\alpha}{x_A}^{\alpha},$$

and it is easy to see that, for a given A, the production function exhibits constant returns to scale; doubling the amount of labor and the amount of each capital good will exactly double output.

It turns out for technical reasons to be easier to analyze the model if we replace the summation in the production function with an integral:

$$Y = {L_Y}^{1-\alpha}\int_0^A {x_j}^{\alpha}\,dj.$$

Then, A measures the range of capital goods that are available to the final-goods sector, and this range is the interval on the real line $[0, A]$. The basic interpretation of this equation, though, is unaffected by this technicality.

With constant returns to scale, the number of firms cannot be pinned down, so we will assume there are a large number of identical firms producing final output and that perfect competition prevails in this sector. We will also normalize the price of the final output, Y, to unity.

Firms in the final-goods sector have to decide how much labor and how much of each capital good to use in producing output. They do this by solving the profit-maximization problem:

$$\max_{L_Y, x_j} {L_Y}^{1-\alpha}\int_0^A {x_j}^{\alpha}\,dj - wL_Y - \int_0^A p_j x_j\,dj,$$

where p_j is the rental price for capital good j and w is the wage paid for labor. The first-order conditions characterizing the solution to this problem are

$$w = (1-\alpha)\frac{Y}{L_Y} \tag{5.12}$$

and

$$p_j = \alpha {L_Y}^{1-\alpha}{x_j}^{\alpha-1}, \tag{5.13}$$

where this second condition applies to each capital good j. The first condition says that firms hire labor until the marginal product of labor equals the wage. The second condition says the same thing, but for capital goods: firms rent capital goods until the marginal product of each kind of capital equals the rental price, p_j. To see the intuition for these equations, suppose the marginal product of a capital good were higher than its rental price. Then the firm should rent another unit; the output produced will more than pay for the rental price. If the marginal product is below the rental price, then the firm can increase profits by reducing the amount of capital used.

5.2.2 THE INTERMEDIATE-GOODS SECTOR

The intermediate-goods sector consists of monopolists who produce the capital goods that are sold to the final-goods sector. These firms gain their monopoly power by purchasing the design for a specific capital good from the research sector. Because of patent protection, only one firm manufactures each capital good.

Once the design for a particular capital good has been purchased (a fixed cost), the intermediate-goods firm produces the capital good with a very simple production function: one unit of raw capital can be automatically translated into one unit of the capital good. The profit maximization problem for an intermediate goods firm is then

$$\max_{x_j} \pi_j = p_j(x_j)x_j - rx_j,$$

where $p_j(x)$ is the demand function for the capital good given in equation (5.13). The first-order condition for this problem, dropping the j subscripts, is

$$p'(x)x + p(x) - r = 0.$$

Rewriting this equation we get

$$p'(x)\frac{x}{p} + 1 = \frac{r}{p},$$

which implies that

$$p = \frac{1}{1 + \frac{p'(x)x}{p}}r.$$

Finally, the elasticity, $p'(x)x/p$, can be calculated from the demand curve in equation (5.13). It is equal to $\alpha - 1$, so the intermediate-goods

firm charges a price that is simply a markup over marginal cost, r:

$$p = \frac{1}{\alpha} r.$$

This is the solution for each monopolist, so that all capital goods sell for the same price. Because the demand functions in equation (5.13) are also the same, each capital good is employed by the final-goods firms in the same amount: $x_j = x$. Therefore, each capital-goods firm earns the same profit. With some algebra, one can show that this profit is given by

$$\pi = \alpha(1 - \alpha)\frac{Y}{A}. \tag{5.14}$$

Finally, the total demand for capital from the intermediate-goods firms must equal the total capital stock in the economy:

$$\int_0^A x_j dj = K.$$

Since the capital goods are each used in the same amount, x, this equation can be used to determine x:

$$x = \frac{K}{A}. \tag{5.15}$$

The final-goods production function can be rewritten, using the fact that $x_j = x$, as

$$Y = AL_Y{}^{1-\alpha}x^{\alpha},$$

and substituting from equation (5.15) reveals that

$$\begin{aligned} Y &= AL_Y{}^{1-\alpha}A^{-\alpha}K^{\alpha} \\ &= K^{\alpha}(AL_Y)^{1-\alpha}. \end{aligned} \tag{5.16}$$

That is, we see that the production technology for the final-goods sector generates the same aggregate production function used throughout this book. In particular, this is the aggregate production function used in equation (5.1).

5.2.3 THE RESEARCH SECTOR

Much of the analysis of the research sector has already been provided. The research sector is essentially like gold mining in the wild West in the mid-nineteenth century. Anyone is free to "prospect" for ideas, and the reward for prospecting is the discovery of a "nugget" that can be sold. Ideas in this model are designs for new capital goods: a faster computer chip, a method for genetically altering corn to make it more resistant to pests, or a new way to organize movie theaters. These designs can be thought of as instructions that explain how to transform a unit of raw capital into a unit of a new capital good. New designs are discovered according to equation (5.4).

When a new design is discovered, the inventor receives a patent from the government for the exclusive right to produce the new capital good. (To simplify the analysis, we assume that the patent lasts forever.) The inventor sells the patent to an intermediate-goods firm and uses the proceeds to consume and save, just like any other agent in the model. But what is the price of a patent for a new design?

We assume that anyone can bid for the patent. How much will a potential bidder be willing to pay? The answer is the present discounted value of the profits to be earned by an intermediate-goods firm. Any less, and someone would be willing to bid higher; any more, and no one would be willing to bid. Let P_A be the price of a new design, this present discounted value. How does P_A change over time? The answer lies in an extremely useful line of reasoning in economics and finance called the method of *arbitrage*.

The arbitrage argument goes as follows. Suppose I have some money to invest for one period. I have two options. First, I can put the money in the "bank" (in this model, this is equivalent to purchasing a unit of capital) and earn the interest rate r. Alternatively, I can purchase a patent for one period, earn the profits that period, and then sell the patent. In equilibrium, it must be the case that the rate of return from both of these investments is the same. If not, everyone would jump at the more profitable investment, driving its return down. Mathematically, the *arbitrage equation* states that the returns are the same:

$$rP_A = \pi + \dot{P}_A. \tag{5.17}$$

The left-hand side of this equation is the interest earned from investing

P_A in the bank; the right-hand side is the profits plus the capital gain or loss that results from the change in the price of the patent. These two must be equal in equilibrium.

Rewriting equation (5.17) slightly,

$$r = \frac{\pi}{P_A} + \frac{\dot{P}_A}{P_A}.$$

Along a balanced growth path, r is constant.[9] Therefore, π/P_A must be constant also, which means that π and P_A have to grow at the same rate; this rate turns out to be the population growth rate, n.[10] Therefore, the arbitrage equation implies that

$$P_A = \frac{\pi}{r - n}. \tag{5.18}$$

This equation gives the price of a patent along a balanced growth path.

5.4 SOLVING THE MODEL

We have now described the market structure and the microeconomics underlying the basic equations given in Section 5.1. The model is somewhat complicated, but several features that were discussed in Chapter 4 are worth noting. First, the aggregate production function exhibits increasing returns. There are constant returns to K and L, but increasing returns once we note that ideas, A, are also an input to production. Second, the increasing returns require imperfect competition. This appears in the model in the intermediate-goods sector. Firms in this sector are monopolists, and capital goods sell at a price that is greater than marginal cost. However, the profits earned by these firms are extracted by the inventors, and these profits simply compensate the inventors for the time they spend "prospecting" for new designs. This framework is called *monopolistic competition*. There are no economic profits in the model; all rents compensate some factor input. Finally, once we depart from the world of perfect competition there is no reason to think that

[9]The interest rate r is constant for the usual reasons. It will be the price at which the supply of capital is equal to the demand for capital, and will be proportional to Y/K.

[10]To see this, recall from equation (5.14) that π is proportional to Y/A. Per capita output, y, and A grow at the same rate, so that Y/A will grow at the rate of population growth.

markets yield the "best of all possible worlds." This last point is one that we develop more carefully in the next section.

We have already solved for the growth rate of the economy in steady state. The part of the model that remains to be solved is the allocation of labor between research and the final-goods sector. Rather than assuming s_R is constant, we let it be determined endogenously by the model.

Once again, the concept of arbitrage enters. It must be the case that, at the margin, individuals in this simplified model are indifferent between working in the final-goods sector and working in the research sector. Labor working in the final-goods sector earns a wage that is equal to its marginal product in that sector, as given in equation (5.12):

$$w_Y = (1 - \alpha)\frac{Y}{L_Y}.$$

Researchers earn a wage based on the value of the designs they discover. We will assume that researchers take their productivity in the research sector, $\bar{\delta}$, as given. They do not recognize that productivity falls as more labor enters because of duplication, and they do not internalize the knowledge spillover associated with ϕ. Therefore, the wage earned by labor in the research sector is equal to its marginal product, $\bar{\delta}$, multiplied by the value of the new ideas created, P_A:

$$w_R = \bar{\delta} P_A.$$

Because there is free entry into both the research sector and the final goods sector, these wages must be the same: $w_Y = w_R$. This condition, with some algebra shown in the appendix to this chapter, reveals that the share of the population that works in the research sector, s_R, is given by

$$s_R = \frac{1}{1 + \frac{r-n}{\alpha g_A}}. \tag{5.19}$$

Notice that the faster the economy grows (the higher is g_A), the higher the fraction of the population that works in research. The higher the discount rate that applies to current profits to get the present discounted value ($r - n$), the lower the fraction working in research.[11]

[11]One can eliminate the interest rate from this equation by noting that $r = \alpha^2 Y/K$ and getting the capital-output ratio from the capital accumulation equation: $Y/K = (n + g + d)/s_K$.

With some algebra, one can show that the interest rate in this economy is given by $r = \alpha^2 Y/K$. Notice that this is *less* than the marginal product of capital, which from equation (5.16) is the familiar $\alpha Y/K$. This difference reflects an important point. In the Solow model with perfect competition and constant returns to scale, all factors are paid their marginal products: $r = \alpha Y/K$, $w = (1 - \alpha)Y/L$, and therefore $rK + wL = Y$. In the Romer model, however, production in the economy is characterized by increasing returns and all factors cannot be paid their marginal products. This is clear from the Solow example just given: because $rK + wL = Y$, there is no output in the Solow economy remaining to compensate individuals for their effort in creating new A. *This* is what necessitates imperfect competition in the model. Here, capital is paid less than its marginal product, and the remainder is used to compensate researchers for the creation of new ideas.

OPTIMAL R&D

Is the share of the population that works in research optimal? In general, the answer to this question in the Romer model is no. In this case, the markets do not induce the right amount of labor to work in research. Why not? Where does Adam Smith's invisible hand go wrong?

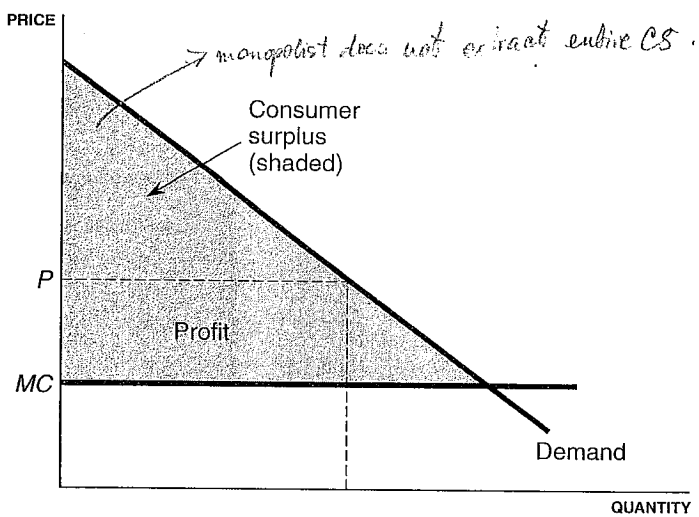
There are three distortions to research in the model that cause s_R to differ from its optimal level. Two of the distortions are easy to see from the production function for ideas. First, the market values research according to the stream of profits that are earned from the new design. What the market misses, though, is that the new invention may affect the productivity of future research. Recall that $\phi > 0$ implies that the productivity of research increases with the stock of ideas. The problem here is one of a missing market: researchers are not compensated for their contribution toward improving the productivity of future researchers. For example, subsequent generations did not reward Isaac Newton sufficiently for inventing calculus. Therefore, with $\phi > 0$, there is a tendency, other things being equal, for the market to provide too little research. This distortion is often called a "knowledge spillover" because some of the knowledge created "spills over" to other researchers. This is the "standing on shoulders" effect referred to earlier. In this sense, it is very much like a classic positive externality: if the bees that

a farmer raises for honey provide an extra benefit to the community that the farmer doesn't capture (they pollinate the apple trees in the surrounding area), the market will underprovide honey bees.[12]

The second distortion, the "stepping on toes" effect, is also a classic externality. It occurs because researchers do not take into account the fact that they lower research productivity through duplication when λ is less than 1. In this case, however, the externality is negative. Therefore, the market tends to provide too much research, other things being equal.

Finally, the third distortion can be called a "consumer-surplus effect." The intuition for this distortion is simple and can be seen by considering a standard monopoly problem, as in Figure 5.4. An inventor of a new design captures the monopoly profit shown in the figure. However, the potential gain to society from inventing the good is the entire consumer-surplus triangle above the marginal cost of production (MC). The incentive to innovate, the monopoly profit, is less than the gain to society, and this effect tends to generate too little innovation, other things being equal.

FIGURE 5.4 THE "CONSUMER-SURPLUS EFFECT"



[12]On the other hand, if $\phi < 0$, then the reverse could be true.

In practice, these distortions can be very large. Consider the consumer surplus associated with basic inventions such as the cure for malaria or cholera or the discovery of calculus. For these inventions, associated with "basic science," the knowledge spillovers and the consumer-surplus effects are generally felt to be so large that the government funds basic research in universities and research centers.

These distortions may also be important even for R&D undertaken by firms. Consider the consumer surplus benefits from the invention of the telephone, electric power, the laser, and the transistor. A substantial literature in economics, led by Zvi Griliches, Edwin Mansfield, and others, seeks to estimate the "social" rate of return to research performed by firms. Griliches (1991) reviews this literature and finds social rates of return on the order of 40 to 60 percent, far exceeding private rates of return. As an empirical matter, this suggests that the positive externalities of research outweigh the negative externalities so that the market, even in the presence of the modern patent system, tends to provide too little research.

A final comment on imperfect competition and monopolies is in order. Classical economic theory argues that monopolies are bad for welfare and efficiency because they create "deadweight losses" in the economy. This reasoning underlies regulations designed to prevent firms from pricing above marginal cost. In contrast, the economics of ideas suggests that it is critical that firms be allowed to price above marginal cost. It is exactly this wedge that provides the profits that are the incentive for firms to innovate. In deciding antitrust issues, modern regulation of imperfect competition has to weigh the deadweight losses against the incentive to innovate.

5.4 SUMMARY

Technological progress is the engine of economic growth. In this chapter, we have endogenized the process by which technological change occurs. Instead of "manna from heaven," technological progress arises as individuals seek out new ideas in an effort to capture some of the social gain these new ideas generate in the form of profit. Better mousetraps get invented and marketed because people will pay a premium for a better way to catch mice.

In Chapter 4, we showed that the nonrivalrous nature of ideas implies that production is characterized by increasing returns to scale. In this chapter, this implication served to illustrate the general importance of scale in the economy. Specifically, the growth rate of world technology is tied to the growth rate of the population. A larger number of researchers can create a larger number of ideas, and it is this general principle that generates per capita growth.

As in the Solow model, comparative statics in this model (such as an increase in the investment rate or an increase in the share of the labor force engaged in R&D) generate *level effects* rather than long-run growth effects. For example, a government subsidy that increases the share of labor in research will typically increase the growth rate of the economy, but only temporarily, as the economy transits to a higher level of income.

The results of this chapter match up nicely with the historical evidence documented in Chapter 4. Consider broadly the history of economic growth in reverse chronological order. The Romer model is clearly meant to describe the evolution of technology since the establishment of intellectual property rights. It is the presence of patents and copyrights that enables inventors to earn profits to cover the initial costs of developing new ideas. In the last century (or two), the world economy has witnessed sustained, rapid growth in population, technology, and per capita income never before seen in history.

Consider how the model economy would behave in the absence of property rights. In this case, innovators would be unable to earn the profits that encourage them to undertake research in the first place, so that no research would take place. With no research, no new ideas would be created, technology would be constant, and there would be no per capita growth in the economy. Broadly speaking, just such a situation prevailed in the world prior to the Industrial Revolution.[13]

Finally, a large body of research suggests that social returns to innovation remain well above private returns. Although the "prizes" that the market offers to potential innovators are substantial, these prizes

[13]There were, of course, very notable scientific and technological advances before 1760, but these were intermittent and there was little sustained growth. What did occur might be attributed to individual curiosity, government rewards, or public funding (such as the prize for the chronometer and the support for astronomical observatories).

still fall far short of the total gain to society from innovations. This gap between social and private returns suggests that large gains are still available from the creation of new mechanisms designed to encourage research. Mechanisms like the patent system are themselves ideas, and there is no reason to think the best ideas have already been discovered.

APPENDIX: SOLVING FOR THE R&D SHARE

The share of the population that works in research, s_R, is obtained by setting the wage in the final-goods sector equal to the wage in research:

$$\bar{\delta} P_A = (1-\alpha)\frac{Y}{L_Y}.$$

Substituting for P_A from equation (5.18),

$$\bar{\delta}\frac{\pi}{r-n} = (1-\alpha)\frac{Y}{L_Y}.$$

Recall that π is proportional to Y/A in equation (5.14):

$$\frac{\bar{\delta}}{r-n}\alpha(1-\alpha)\frac{Y}{A} = (1-\alpha)\frac{Y}{L_Y}.$$

Several terms cancel, leaving

$$\frac{\alpha}{r-n}\frac{\bar{\delta}}{A} = \frac{1}{L_Y}.$$

Finally, notice that $\dot{A}/A = \bar{\delta} L_A/A$, so that $\bar{\delta}/A = g_A/L_A$ along a balanced growth path. With this substitution,

$$\frac{\alpha g_A}{r-n} = \frac{L_A}{L_Y}.$$

Notice that L_A/L_Y is just $s_R/(1-s_R)$. Solving the equation for s_R then reveals

$$s_R = \frac{1}{1+\frac{r-n}{\alpha g_A}},$$

as reported in equation (5.19).

EXERCISES

1. *An increase in the productivity of research.* Suppose there is a one-time increase in the productivity of research, represented by an increase in $\bar{\delta}$ in Figure 5.1. What happens to the growth rate and the level of technology over time?

2. *Too much of a good thing?* Consider the level of per capita income along a balanced growth path given by equation (5.11). Find the value for s_R that maximizes output per worker along a balanced growth path for this example. Why is it possible to do too much R&D according to this criterion?

3. *The future of economic growth* (from Jones (2002)). Recall from Figure 4.6 and the discussion surrounding this figure in Chapter 4 that the number of scientists and engineers engaged in R&D has been growing faster than the rate of population growth in the advanced economies of the world. To take some plausible numbers, assume population growth is 1 percent and the growth rate of researchers is 3 percent per year. Assume that $\dot{A}/A$ has been constant at about 2 percent per year.

   (a) Using equation (5.6), calculate an estimate of $\lambda/(1-\phi)$.

   (b) Using this estimate and equation (5.7), calculate an estimate of the long-run steady-state growth rate of the world economy.

   (c) Why does your estimate of long-run steady-state growth differ from the 2 percent rate of growth of A observed historically?

   (d) Does the fact that many developing countries are starting to engage in R&D change this calculation?

4. *The share of the surplus appropriated by inventors* (from Kremer 1998). In Figure 5.4, find the ratio of the profit captured by the monopolist to the total potential consumer surplus available if the good were priced at marginal cost. Assume that marginal cost is constant at c and the demand curve is linear: $Q = a - bP$, where a, b, and c are positive constants with $a - bc > 0$.

6 A SIMPLE MODEL OF GROWTH AND DEVELOPMENT

The neoclassical growth model allows us to think about why some countries are rich while others are poor, taking technology and factor accumulation as exogenous. The Romer model provides the microeconomic underpinnings for a model of the technological frontier and why technology grows over time. It answers in detail our questions concerning the "engine of growth." In this chapter, we address the next logical question, which is how technologies diffuse across countries, and why the technology used in some countries is so much more advanced than the technology used in others.

6.1 THE BASIC MODEL

The framework we develop builds naturally on the Romer model of technology discussed in Chapter 5. The component that we add to the model is an avenue for technology transfer. We endogenize the mechanism by which different countries achieve the ability to use various intermediate capital goods.

As with the Romer model, countries produce a homogeneous output good, Y, using labor, L, and a range of capital goods, x_j. The "number" of capital goods that workers can use is limited by their skill level, h:[1]

$$Y = L^{1-\alpha} \int_0^h x_j^{\alpha} dj. \tag{6.1}$$

Once again, think of the integral as a sum. A worker with a high skill level can use more capital goods than a worker with a low skill level. For example, a highly skilled worker may be able to use computerized machine tools unavailable to workers below a certain skill level.

In Chapter 5, we focused on the invention of new capital goods as an engine of growth for the world economy. Here, we will have the opposite focus. We assume that we are examining the economic performance of a single small country, potentially far removed from the technological frontier. This country grows by learning to use the more advanced capital goods that are already available in the rest of the world. Whereas the model in Chapter 5 can be thought of as applying to the OECD or the world as a whole, this model is best applied to a specific economy.

One unit of any intermediate capital good can be produced with one unit of raw capital. To simplify the setup, we assume this transformation is effortless and can also be undone effortlessly. Thus,

$$\int_0^{h(t)} x_j(t)\, dj = K(t), \tag{6.2}$$

that is, the total quantity of capital goods of all types used in production is equal to the total supply of raw capital. Intermediate goods are treated symmetrically throughout the model, so that $x_j = x$ for all j. This fact, together with equation (6.2) and the production function in (6.1), implies that the aggregate production technology for this economy takes the familiar Cobb-Douglas form:

$$Y = K^{\alpha}(hL)^{1-\alpha}. \tag{6.3}$$

Notice that an individual's skill level, h, enters the equation just like labor-augmenting technology.

[1]This production function is also considered by Easterly, King, et al. (1994).

Capital, K, is accumulated by forgoing consumption, and the capital accumulation equation is standard:

$$\dot{K} = s_K Y - dK,$$

where s_K is the investment share of output (the rest going to consumption) and d is some constant exponential rate of depreciation greater than 0.

Our model differs from that in Chapter 3 in terms of the accumulation of skill h. There, an individual's skill level was simply a function of the amount of time the individual spent in school. Here, we generalize this idea as follows. "Skill" is now defined specifically as the range of intermediate goods that an individual has learned to use. As individuals progress from using hoes and oxen to using pesticides and tractors, the economy grows. Individuals learn to use more advanced capital goods according to

$$\dot{h} = \mu e^{\psi u} A^{\gamma} h^{1-\gamma}. \tag{6.4}$$

In this equation, u denotes the amount of time an individual spends accumulating skill instead of working. Empirically, we might think of u as years of schooling, although clearly individuals also learn skills outside of formal education. A denotes the world technology frontier. It is the index of the most advanced capital good invented to date. We assume $\mu > 0$ and $0 < \gamma \leq 1$.[2]

Equation (6.4) has a number of features that merit discussion. First, notice that we preserve the basic exponential structure of skill accumulation. Spending additional time accumulating skill will increase the skill level proportionally. As in Chapter 3, this is intended to match the microeconomic evidence on returns to schooling. Second, the last two terms suggest that the change in skill is a (geometrically) weighted average of the frontier skill level, A, and the individual's skill level, h.

To see more clearly what equation (6.4) implies about skill accumulation, it can be rewritten by dividing both sides by h:

$$\frac{\dot{h}}{h} = \mu e^{\psi u} \left(\frac{A}{h}\right)^{\gamma}. \tag{6.5}$$

[2]Equation (6.4) is reminiscent of a relationship analyzed by Nelson and Phelps (1966) and more recently by Bils and Klenow (2000).

This equation makes clear the implicit assumption that it is harder to learn to use an intermediate good that is currently close to the frontier. The closer an individual's skill level, h, is to the frontier, A, the smaller the ratio A/h, and the slower his or her skill accumulation. This implies, for example, that it took much longer to learn to use computers thirty years ago, when they were very new, than it does today.

The technological frontier is assumed to evolve because of investment in research by the advanced economies in the world. Drawing on the results of the Romer model, we assume that the technology frontier expands at a constant rate, g:

$$\frac{\dot{A}}{A} = g.$$

A more complete model would allow individuals to choose to work in either the final-goods sector or in research, along the lines of Chapter 5. In a model like this, g would be a function of the parameters of the production function for ideas and the world rate of population growth. To simplify the analysis, however, we will not develop this more complete story. In this model, we assume that there is a world pool of ideas that are freely available to any country. In order to take advantage of these ideas, however, a country must first learn to use them.

6.2 STEADY-STATE ANALYSIS

As in earlier chapters, we will assume that the investment rate in the economy and the amount of time individuals spend accumulating skill instead of working are given exogenously and are constant. This is increasingly becoming an unpleasant assumption, and it is one we will explore in much greater detail in Chapter 7. We also assume the labor force of the economy grows at the constant, exogenous rate n.

To solve for the balanced growth path in this economy, consider the skill accumulation equation in (6.5). Along a balanced growth path, the growth rate of h must be constant. Recall that since h enters the production function in equation (6.3) just like labor-augmenting technology, the growth rate of h will pin down the growth rate of output per worker, $y \equiv Y/L$, and capital per worker, $k \equiv K/L$, as well. From equation (6.5), we see that $\dot{h}/h$ will be constant if and only if A/h is constant, so that

h and A must grow at the same rate. Therefore, we have

$$g_y = g_k = g_h = g_A = g, \tag{6.6}$$

where as usual g_x denotes the growth rate of the variable x. The growth rate of the economy is given by the growth rate of human capital or skill, and this growth rate is tied down by the growth rate of the world technological frontier.

To solve for the level of income along this balanced growth path, we proceed in the usual fashion. The capital accumulation equation implies that along a balanced growth path the capital-output ratio is given by

$$\left(\frac{K}{Y}\right)^* = \frac{s_K}{n+g+d}.$$

Substituting this into the production function in equation (6.3) after rewriting it in terms of output per worker, we have

$$y^*(t) = \left(\frac{s_K}{n+g+d}\right)^{\alpha/1-\alpha} h^*(t), \tag{6.7}$$

where the asterisk (*) is used to indicate variables along a balanced growth path. We have made explicit the fact that y and h are changing over time by including the t index.

Along the balanced growth path, the ratio of the skill level in our small economy to the most advanced capital good invented to date is pinned down by the accumulation equation for skill, equation (6.5). Using the fact that $g_h = g$, we know that

$$\left(\frac{h}{A}\right)^* = \left(\frac{\mu}{g}e^{\psi u}\right)^{1/\gamma}.$$

This equation tells us that the more time individuals spend accumulating skills, the closer the economy is to the technological frontier.[3]

Using this equation to substitute for h in equation (6.7), we can write output per worker along the balanced growth path as a function

[3]To be sure that the ratio h/A is less than one, we assume μ is sufficiently small.

of exogenous variables and parameters:

$$y^*(t) = \left(\frac{s_K}{n+g+d}\right)^{\alpha/1-\alpha} \left(\frac{\mu}{g}e^{\psi u}\right)^{1/\gamma} A^*(t). \tag{6.8}$$

Equations (6.6) and (6.8) represent the key equations that describe the implications of our simple model for economic growth and development. Recall that equation (6.6) states that along a balanced growth path, output per worker increases at the rate of the skill level of the labor force. This growth rate is given by the growth rate of the technological frontier.

Equation (6.8) characterizes the level of output per worker along this balanced growth path. The careful reader will note the similarity between this equation and the solution of the neoclassical model in equation (3.8) of Chapter 3. The model developed in this chapter, emphasizing the importance of ideas and technology transfer, provides a "new growth theory" interpretation of the basic neoclassical growth model. Here, economies grow because they learn to use new ideas invented throughout the world.

Several other remarks concerning this equation are in order. First, the initial term in equation (6.8) is familiar from the original Solow model. This term says that economies that invest more in physical capital will be richer, and economies that have rapidly growing populations will be poorer.

The second term in equation (6.8) reflects the accumulation of skills. Economies that spend more time accumulating skills will be closer to the technological frontier and will be richer. Notice that this term is similar to the human capital term in our extension of the Solow model in Chapter 3. However, now we have made explicit what the accumulation of skill means. In this model, skills correspond to the ability to use more advanced capital goods. As in Chapter 3, the way skill accumulation affects the determination of output is consistent with microeconomic evidence on human capital accumulation.

Third, the last term of the equation is simply the world technological frontier. This is the term that generates the growth in output per worker over time. As in earlier chapters, the engine of growth in this model is technological change. The difference relative to Chapter 3 is that we now understand from the analysis of the Romer model where technological change comes from.

Fourth, the model proposes one answer to the question of why different economies have different levels of technology. Why is it that high-tech machinery and new fertilizers are used in producing agricultural products in the United States while agriculture in India or sub-Saharan Africa relies much more on labor-intensive techniques? The answer emphasized by this model is that the skill level of individuals in the United States is much higher than the skill level of individuals in developing countries. Individuals in developed countries have learned over the years to use very advanced capital goods, while individuals in developing countries have invested less time in learning to use these new technologies.

Implicit in this explanation is the assumption that technologies are available worldwide for anyone to use. At some level, this must be a valid assumption. Multinational corporations are always looking around the world for new places to invest, and this investment may well involve the use of advanced technologies. For example, cellular phone technology has proved very useful in an economy such as China's: instead of building the infrastructure associated with telephone lines and wires, several companies are vying to provide cellular communications. Multinational companies have signed contracts to build electric power grids and generators in a number of countries, including India and the Philippines. These examples suggest that technologies may be available to flow very quickly around the world, provided the economy has the infrastructure and training to use the new technologies.

By explaining differences in technology with differences in skill, this model cannot explain one of the empirical observations made in Chapter 3. There, we calculated total factor productivity (TFP)—the productivity of a country's inputs, including physical and human capital, taken together—and documented that TFP levels vary considerably across countries. This variation is not explained by the model at hand, in which all countries have the same level of total factor productivity. What then explains the differences? This is one of the questions we address in the next chapter.[4]

[4]Strictly speaking, we must be careful in applying the evidence from Chapter 3 to this model. For example, here the exponent $(1/\gamma)$ on time spent accumulating skills is an additional parameter.

6.3 TECHNOLOGY TRANSFER

In the model we have just outlined, technology transfer occurs because individuals in an economy learn to use more advanced capital goods. To simplify the model, we assumed that the designs for new capital goods were freely available to the intermediate-goods producers.

The transfer of technology is likely to be more complicated than this in practice. For example, one could imagine that the designs for new capital goods have to be altered slightly in different countries. The steering wheel on an automobile may need to be switched to the other side of the car, or the power source for an electrical device may need to be altered to conform to a different standard.

Technology transfer also raises the issue of international patent protection. Are intellectual property rights assigned in one country enforced in another? If so, then new designs may need to be licensed from the inventor before they can be used. As noted in Chapter 4, the ability to sell one's ideas in a global marketplace raises the returns to invention, thereby encouraging research.

Costs of adapting or licensing new designs are similar in some ways to the fixed costs of invention. Consider the case in which the inventor of our hypothetical WordTalk software program is deciding whether or not to produce a version of the software for China. Adapting the program to the Chinese language is somewhat like inventing an entirely new program. Substantial up-front development costs may be required to alter the program. The fact that China is a potentially enormous market may make these costs worth paying. But only, of course, if the intellectual property right is respected. In addition, the skills of the Chinese workforce are clearly relevant; it is not simply the number of people in China that matters, but the number of people who have computers and the skills to use them.[5]

[5]This setup is somewhat related to the idea of Basu and Weil (1998) that certain technologies may be appropriate only once a certain level of development has been reached. To use one of their examples, the latest "maglev" trains from Japan may not be useful in an economy such as that of Bangladesh, which depends on bicycles and bullock carts.

6.4 UNDERSTANDING DIFFERENCES IN GROWTH RATES

A key implication of equation (6.8) is that all countries share the same long-run growth rate, given by the rate at which the world technological frontier expands. In Chapters 2 and 3, we simply assumed this result. The simple model of technology transfer developed in this chapter provides one justification for this assumption.[6]

In models based on the diffusion of technology, the outcome that all countries share a common growth rate is typical. Belgium and Singapore do not grow solely or even mainly because of ideas invented by Belgians and Singaporeans. The populations of these economies are simply too small to produce a large number of ideas. Instead, these economies grow over time because they — to a greater or lesser extent — are successful at learning to use new technologies invented throughout the world. The eventual diffusion of technologies, even if it takes a very long time, prevents any economy from falling too far behind.[7]

How does this prediction that all countries have the same long-run growth rate match up with the empirical evidence? In particular, we know that average growth rates computed over two or three decades vary enormously across countries (see Chapter 1). While the U.S. economy grew at 1.4 percent, the Japanese economy grew at 5 percent per year from 1950 to 1990. Differences also exist over very long periods of time. For example, from 1870 to 1994, the United States grew at an average rate of 1.8 percent while the United Kingdom grew much more slowly at 1.3 percent. Doesn't the large variation in average growth rates that we observe empirically contradict this model?

The answer is no, and it is important to understand why. The reason is the one we have already discussed in Chapter 3. Even with no difference across countries in the long-run growth rate, we can explain the large variation in rates of growth with *transition dynamics*. To the extent that countries are changing their position within the long-run income

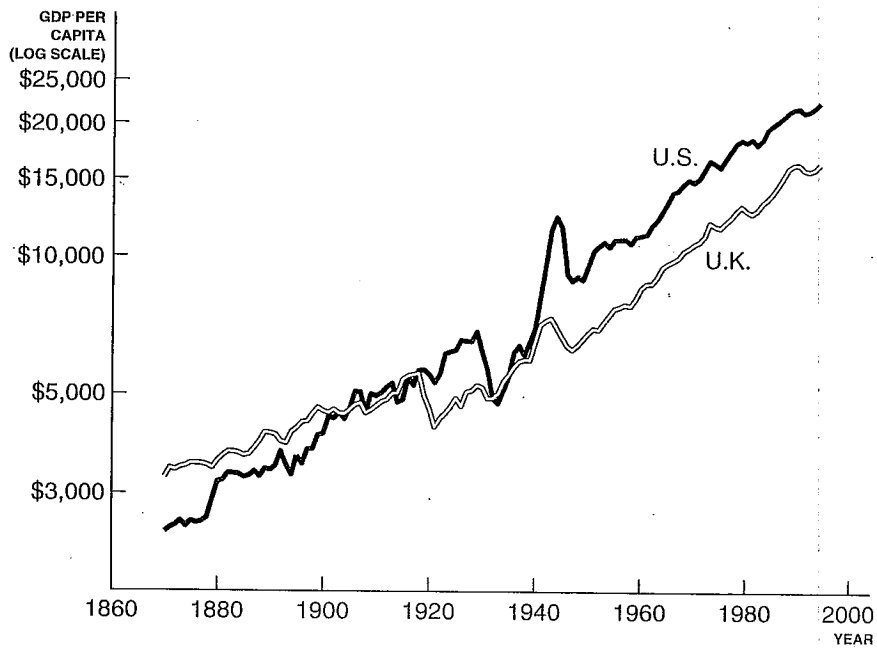
[6]The remainder of this section draws on Jones (1997).

[7]One important exception is notable and will be discussed further in Chapter 7. Suppose that policies in an economy are so bad that individuals are not allowed to earn a return on their investments. This may prevent anyone from investing at all, which may result in a "development trap" in which the economy does not grow.

distribution, they can grow at different rates. Countries that are "below" their steady-state balanced growth paths should grow faster than g, and countries that are "above" their steady-state balanced growth paths should grow more slowly. What causes an economy to be away from its steady state? Any number of things. A shock to the country's capital stock (e.g., it is destroyed in a war) is a typical example. A policy reform that increases the investment in capital and skill accumulation is another.

This general point can be illustrated by taking a closer look at the behavior of the United States and the United Kingdom over the last 125 years. Figure 6.1 plots the log of GDP per capita in these two countries from 1870 to 1994. As noted above, growth in the United States over this period was fully a half-point higher than it was in the United Kingdom. However, a careful look at Figure 6.1 reveals that nearly all of this difference occurred before 1950, as the United States overtook

FIGURE 6.1 INCOME IN THE U.S. AND THE U.K., 1870–1994



SOURCE: Maddison (1995).

the United Kingdom as the world's leading economy. From 1870 to 1950, the United States grew at 1.7 percent per year while the United Kingdom grew at only 0.9 percent. Since 1950, however, growth in these two economies has been nearly identical. The United States grew at 1.95 percent per year from 1950 to 1994 while the United Kingdom grew at 1.98 percent.

This example suggests that we have to be very careful in interpreting differences in average growth rates across economies. Even over very long periods of time they may differ. This is exactly what our model predicts. However, this does not mean that the underlying long-run growth rate varies across economies. The fact that Japan has grown much faster than the United States over the last forty years tells us very little about the underlying long-run growth rate of these economies. To infer that Japan will continue its astounding performance would be analogous to concluding in 1950 that growth in the United States would be permanently higher than growth in the United Kingdom. History has shown us that this second inference, at least, was incorrect.

The model in this chapter illustrates another important point. The principle of transition dynamics is not simply a feature of the capital accumulation equation in the neoclassical growth model, as was the case in Chapter 3. In this model, transition dynamics involve not only capital accumulation but also the technology transfer specification in equation (6.4). For example, suppose a country decides to reduce tariffs and trade barriers and open up its economy to the rest of the world. This policy reform might enhance the ability of the economy to transfer technologies from abroad; we can model this as an increase in μ. According to equation (6.8), a higher value of μ raises the economy's steady-state level of income. This means that at its current level, the economy is now below its steady-state income. What happens when this is the case? The principle of transition dynamics tells us that the economy grows rapidly as it transits to the higher income level.

EXERCISES

1. *The importance of* A *versus* h *in producing human capital.* How might one pick a value of γ to be used in the empirical analysis of the model (as in Chapter 3)? Other things equal, use this value to discuss how differences in skills affect output per worker in the steady state, compared to the model used in Chapter 3.

2. *Understanding levels of income.* This model explains differences in the level of income across countries by appealing to differences in s_K and u. What is unsatisfying about this explanation?

3. *Understanding growth rates.* How does the model explain the differences in growth rates that we observe across countries?

4. *The role of* μ. Provide some economic intuition for the role played by the parameter μ. What values of μ guarantee that h/A is less than 1?

5. *Openness to technology transfer.* This problem considers the effect on an economy's technological sophistication of an increase in the openness of the economy to technology transfer. Specifically, it looks at the short-run and long-run effects on h of an increase in μ. (Hint: Look back at Figure 5.1 in Chapter 5.)

   (a) Construct a graph with $\dot{h}/h$ on the vertical axis and A/h on the horizontal axis. In the graph, plot two lines:

   $$\frac{\dot{h}}{h} = \mu e^{\psi u}\left(\frac{A}{h}\right)$$

   and

   $$\dot{h}/h = g.$$

   (Note that we've assumed $\gamma = 1$.) What do these two lines mean, and what is the significance of the point where they intersect?

   (b) Starting from steady state, analyze the short-run and long-run effects of an increase in μ on the growth rate of h.

   (c) Plot the behavior of h/A over time.

   (d) Plot the behavior of $h(t)$ over time (on a graph with a log scale).

   (e) Discuss the consequences of an increase in openness to technology transfer on an economy's technological sophistication.

7 SOCIAL INFRASTRUCTURE AND LONG-RUN ECONOMIC PERFORMANCE

> It is often assumed that an economy of private enterprise has an automatic bias towards innovation, but this is not so. It has a bias only towards profit.
>
> —ERIC J. HOBSBAWM (1969), cited by Baumol (1990), p. 893.

An important assumption maintained by all of the models considered up until now is that the investment rates and the time individuals spend accumulating skill are given exogenously. When we ask why some countries are rich while others are poor, our answer has been that rich countries invest more in capital and spend more time learning to use new technologies. However, this answer begs new questions: Why is it that some countries invest more than others, and why do individuals in some countries spend more time learning to use new technologies?

Addressing these questions is currently one of the important subjects of research by economists who study growth and development, yet no consensus has emerged regarding the answer. As a result, there is no "canonical" model to help us outline an answer, as the Solow and Romer models did for earlier questions. Nevertheless, theory is such a useful way to organize one's thoughts that this chapter will present a very basic framework for thinking about these questions. The framework is

motivated by a simple investment problem of the kind faced by business managers every day.[1]

7.1 A BUSINESS INVESTMENT PROBLEM

Suppose you are the manager of a large, successful multinational corporation, and you are considering opening a subsidiary in a foreign country. How do you decide whether to undertake this investment?

One approach to evaluating this investment project is called *cost-benefit analysis*. In such an analysis, we calculate the total costs of the project and the total benefits, and if the benefits are larger, then we proceed.

Suppose that launching the business subsidiary involves a one-time setup cost F. For example, establishing the subsidiary may require obtaining both domestic and foreign business licenses, as well as business contacts with suppliers and distributors in the foreign country.

Once the business is set up, let's assume that it generates a profit every year that the business remains open. If Π denotes the expected present discounted value of the profit stream, then Π is the value of the business subsidiary once it has been set up. Why? Suppose that the parent company decides to sell the subsidiary after the one-time setup cost F has been paid. How much would another company be willing to pay to purchase the subsidiary? The answer is the present discounted value of the future profits, or at least what we expect them to be. This is exactly Π.

With this basic formalization of the investment problem, deciding whether or not to undertake the project is straightforward. If the value of the business after it is set up is larger than the cost of setting up the subsidiary, then the manager should undertake the investment project. The manager's decision is

$$\Pi \geq F \rightarrow \text{Invest},$$

$$\Pi < F \rightarrow \text{Do not invest}.$$

[1]This chapter expands on a number of ideas presented by Hall and Jones (1999).

Although we have chosen a business project as the example to explain this analysis, the basic framework can be applied to the determination of domestic investment by a local business, the transfer of technology by a multinational corporation, or the decision to accumulate skills by an individual. The extension to technology transfer is inherent in the business example. A substantial amount of technology transfer presumably occurs in exactly this way — when multinational corporations decide to set up a new kind of business in a foreign country. With respect to skill acquisition, a similar story applies. Individuals must decide how much time to spend acquiring specific skills. For example, consider the decision of whether or not to spend another year in school. F is the cost of schooling, both in terms of direct expenditures and in terms of opportunity cost (individuals could spend the time working instead of going to school). The benefit Π reflects the present value of the increase in wages that results from the additional skill acquisition.

What determines the magnitudes of F and Π in various economies around the world? Is there sufficient variation in F and Π to explain the enormous variation in investment rates, educational attainment, and total factor productivity? The hypothesis we will pursue in this chapter is that there is a great deal of variation in the costs of setting up a business and in the ability of investors to reap returns from their investments. Such variation arises in large part from differences in government policies and institutions — what we might call *social infrastructure*. A good government provides the institutions and social infrastructure that minimize F and maximize Π (or, more correctly, maximize $\Pi - F$), thereby encouraging investment.

7.2 DETERMINANTS OF *F*

First, consider the cost of setting up a business subsidiary, F. Establishing a business, even once the *idea* driving the business has been created — say the next "killer application" in computer software, or even the notion that a particular location on a particular street would be a great place to set up a hot-dog stand — requires a number of steps. Each of these steps involves interacting with another party, and if that party has the ability to "hold up" the business, problems can arise. For

example, in setting up a hot-dog stand, the property has to be purchased, the hot-dog stand itself must be inspected by officials, and a business permit may be necessary. Obtaining electricity may require another permit. Each of these steps offers an opportunity for a crafty bureaucrat to seek a bribe or for the government to charge a licensing fee.

These kinds of concerns can be serious. For example, after the land and equipment have been purchased and several permits obtained, what prevents the next bureaucrat — perhaps the one from whom the final license must be obtained — from asking for a bribe equal to Π (or slightly smaller)? At this point, the rational manager, with no choice other than canceling the project, may well be forced to give in and pay the bribe. All of the other fees and bribes that have been paid are "sunk costs" and do not enter the calculation of whether the next fee should be paid.

But, of course, the astute manager will envision this scenario from the very beginning, before any land or equipment is purchased and before any fees and bribes have been paid. The rational choice at this *ex ante* point is not to invest at all.

To residents of advanced countries such as the United States or the United Kingdom, this issue may seem unimportant as a matter of practice. But this, as we will see, is exactly the point. Advanced countries provide a dynamic business environment, full of investment and entrepreneurial talent, exactly *because* such concerns are minimal.

There is a wealth of anecdotal evidence from other countries to suggest that this kind of problem can be quite serious. Consider the following example, which describes the problem of foreign investment in post-Communist Russia:

> To invest in a Russian company, a foreigner must bribe every agency involved in foreign investment, including the foreign investment office, the relevant industrial ministry, the finance ministry, the executive branch of the local government, the legislative branch, the central bank, the state property bureau, and so on. The obvious result is that foreigners do not invest in Russia. Such competing bureaucracies, each of which can stop a project from proceeding, hamper investment and growth around the world, but especially in countries with weak governments (Shleifer and Vishny 1993, pp. 615–16).

Another excellent example of the impact of government policies and institutions on the costs of setting up a business is provided by Her-

nando de Soto's *The Other Path* (1989). Like his more famous namesake, this contemporary de Soto gained renown by opposing the Peruvian establishment. What he sought, however, was not the riches of Peru, but rather the reason for the lack of riches in that country.[2]

In the summer of 1983, de Soto and a team of researchers started a small garment factory on the outskirts of Lima, Peru, for the express purpose of measuring the cost of complying with the regulations, red tape, and bureaucratic restrictions associated with a small entrepreneur starting a business. The researchers were confronted with 11 official requirements, such as obtaining a zoning certificate, registering with the tax authority, and procuring a municipal license. Meeting these official requirements took 289 person-days. Including the payment of 2 bribes (although 10 bribes were requested, "only" 2 were paid because they were absolutely required in order to continue the project), the cost of starting a small business was estimated to be the equivalent of 32 times the monthly minimum living wage.[3]

DETERMINANTS OF Π

Apart from the costs of setting up a business, what are the determinants of the expected profitability of the investment? We will classify these determinants into three categories: (1) the size of the market, (2) the extent to which the economy favors production instead of diversion, and (3) the stability of the economic environment.

The size of the market is one of the critical determinants of Π and therefore one of the critical factors in determining whether or not investments get undertaken. Consider, for example, the development of the Windows XP operating system by Microsoft. Would it have been worth the hundreds of millions of dollars required to develop this program if Microsoft could sell the software only in Washington state? Probably not. Even if every computer in Washington ran the Windows operating system, the revenue from sales of Windows XP would not cover development costs — there are simply too few computers in the state. In

[2]Long before exploring the Mississippi River and the southeastern United States, the more famous Hernando de Soto obtained his wealth as a Spanish conquistador of Peru.
[3]See de Soto (1989).

reality, the market for this software is, quite literally, the world, and the presence of a large market increases the potential reward for making the investment. This is another example of the "scale effect" associated with fixed or one-time costs.

This example suggests another point that is important: the relevant market for a particular investment need not be limited by national borders. The extent to which an economy is open to international trade has a potentially profound influence on the size of the market. For example, building a factory to manufacture hard-disk drives in Singapore may not seem like a good idea if Singapore is the entire market; more people live in the San Francisco Bay area than in the entire country of Singapore. However, Singapore is a natural harbor along international shipping routes and has one of the world's most open economies. From Singapore, one can sell disk drives to the rest of the world.

A second important determinant of the profits to be earned on an investment is the extent to which the rules and institutions in an economy favor *production* or *diversion*. Production needs little explanation: a social infrastructure that favors production encourages individuals to engage in the creation and transaction of goods and services. In contrast, diversion takes the form of the theft or expropriation of resources from productive units. Diversion may correspond to illegal activity, such as theft, corruption, or the payment of "protection money," or it may be legal, as in the case of confiscatory taxation by the government, frivolous litigation, or the lobbying of the government by special interests.

The first effect of diversion on a business is that it acts like a tax. Some fraction of the revenue or profits earned on an investment are taken away from the entrepreneur, detracting from the return on the investment. The second effect of diversion is that it encourages investment by the entrepreneur in finding ways to avoid the diversion. For example, the business may have to hire extra security guards or accountants and lawyers or pay bribes in order to avoid other forms of diversion. Of course, these investments in avoidance are also a form of diversion.

The extent to which the economic environment of a country favors production or diversion is primarily determined by the government. The government makes and enforces the laws that provide the framework for economic transactions in the economy. Moreover, in economies with environments that favor diversion, the government is itself

often a chief agent of diversion. Taxation is a form of diversion, and although some taxation is necessary in order for the government to be able to provide the rules and institutions associated with an infrastructure that favors production, the power to tax can be abused. Red tape and bureaucratic regulation enable government officials to use their influence to divert resources.

The power to make and enforce laws conveys an enormous power to the government to engage in diversion. This suggests the importance of an effective system of checks and balances and the separation of powers among several branches of government. This issue is reminiscent of the well-known aphorism "But who guards the guardians?" attributed to the Roman satirist Juvenal.[4]

Finally, the stability of the economic environment can itself be an important determinant of the returns to investing. An economy in which the rules and institutions are changing frequently may be a risky place in which to invest. Although the policies in place today may favor productive activities in an open economy, perhaps the policies tomorrow will not. Wars and revolutions in an economy are extreme forms of instability.

7.4 WHICH INVESTMENTS TO MAKE?

The institutions and policies of an economy potentially have a large influence on investment. Economies in which the social infrastructure encourages diversion instead of production will typically have less investment in capital, less foreign investment that might transfer technology, less investment by individuals in accumulating productive skills, and less investment by entrepreneurs in developing new ideas that improve the production possibilities of the economy.

In addition, the social infrastructure of the economy may influence the type of investments that are undertaken. For example, in an economy in which theft is a serious problem, managers may invest capital in fences and security systems instead of productive machines and factories. Or in an economy in which government jobs provide the ability to earn rents by collecting taxes or bribes, individuals may invest in accumulating skills that allow them to obtain government employment instead of skills that would enhance production.

[4]Plato, another great writer about guardians, seems to think less of this problem in *The Republic*: "That they must abstain from intoxication has already been remarked by us; for of all persons, a guardian should be the last to get drunk and not know where he is. Yes, he said; that a guardian should require another guardian to take care of him is ridiculous indeed."

7.5 EMPIRICAL EVIDENCE

Our simple theoretical framework for analyzing investments has a number of general predictions. A country that attracts investments in the form of capital for businesses, technology transfer from abroad, and skills from individuals will be one in which

- the institutions and laws favor production over diversion,
- the economy is open to international trade and competition in the global marketplace, and
- the economic institutions are stable.

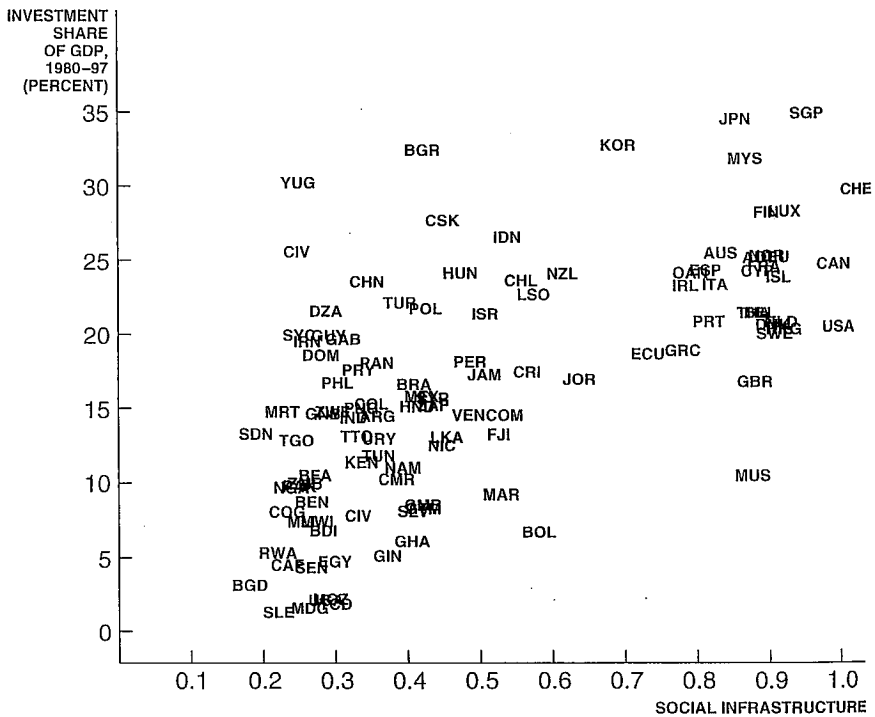
A good social infrastructure encourages domestic investment by firms in physical capital (factories and machines), investment by foreign entrepreneurs that may involve the transfer of better technologies, and the accumulation of skills by individuals. Furthermore, such an environment encourages domestic entrepreneurship; individuals look for better ways to create, produce, or transport their goods and services instead of looking for more effective ways to divert resources from other agents in the economy.

What empirical evidence supports these claims? Ideally, one would like empirical measures of the attributes of an economy that encourage the various forms of investment. Then, one could look at the economies of the world to see if these attributes are associated with high rates of investment and successful economic performance.

Several measures of these attributes are available from a large research literature examining long-run economic performance. Two of these measures have been averaged by Hall and Jones (1999) to create

an index of social infrastructure.[5] The first measure is an index of "government anti-diversion policies" that captures the extent to which the social infrastructure of an economy favors production over diversion. This measure is assembled from a consulting firm that specializes in providing advice to multinational investors. The second variable represents the fraction of years since 1950 that the economy is classified as open to international trade according to several objective criteria. The index is normalized so that a value of 1 represents the best existing infrastructure and a value of 0 represents the worst.

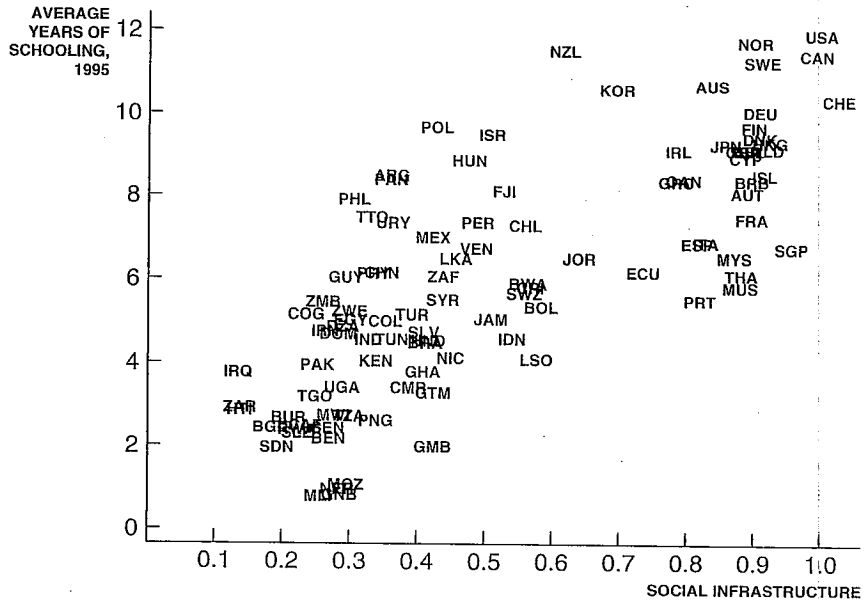
FIGURE 7.1 UNDERSTANDING DIFFERENCES IN INVESTMENT RATES



SOURCE: Author's calculation using data from Appendix C and Hall and Jones (1999).

[5]The construction of this index is discussed in more detail in that paper. Briefly, the underlying data are taken from Knack and Keefer (1995) and Sachs and Warner (1995).

FIGURE 7.2 UNDERSTANDING DIFFERENCES IN SKILL ACCUMULATION



SOURCE: Author's calculation using data from Appendix C and Hall and Jones (1999).

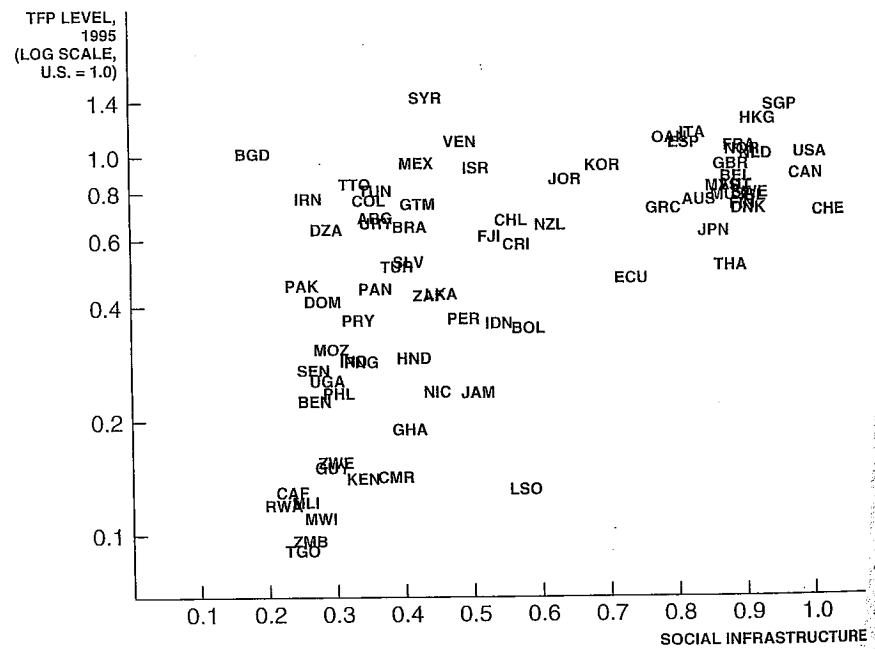
Figures 7.1 and 7.2 plot investment as a share of GDP and average educational attainment against this index of social infrastructure. These figures show a strong relationship between social infrastructure and factor accumulation: countries with a good social infrastructure tend to have much higher investment rates in both physical and human capital. In countries where the social infrastructure allows investors to earn appropriate returns on their investments, firms and workers invest heavily in capital and skills.

This reasoning suggests a possible explanation of the stylized fact related to migration that we discussed in Chapter 1 (Fact 7). Recall that standard neoclassical theory suggests that rates of return are directly related to scarcity. If skilled labor is a scarce factor in developing economies, the return to skill in these economies should be high, and this should encourage the migration of skilled labor out of rich countries and into poor countries. Empirically, however, the opposite pattern seems

to occur. The explanation suggested here reverses this reasoning. Suppose that, at least to a first approximation, rates of return to skill are equalized by migration across countries. The stock of skills in developing countries is so low because skilled individuals are not allowed to earn the full return on their skills. Much of their skill is wasted by diversion — such as the payment of bribes and the risk that the fruits of their skill will be expropriated.[6]

Finally, Figure 7.3 plots the total factor productivity (TFP) level against social infrastructure. Recall from Chapter 3 that some countries get much more output from their inputs (capital and skills) than do

FIGURE 7.3 **UNDERSTANDING DIFFERENCES IN TOTAL FACTOR PRODUCTIVITY**



SOURCE: Author's calculation using data from Appendix C and Hall and Jones (1999).

[6]Migration restrictions could then explain the observed pattern that skilled labor migrates from developing countries to developed countries when it has the opportunity.

other countries. This is reflected in differences in TFP across countries. Figure 7.3 shows that these differences are also related to social infrastructure. To see why this might be the case, consider a simple example in which individuals can choose to be either farmers or thieves. In the economy of Cornucopia, government policies strongly support production, no one is a thief, and society gets the maximum amount of output from its resources. On the other hand, in the economy of Kleptocopia, whose policies do not support production, thievery is an attractive alternative. Some individuals spend their time stealing from farmers. Thus some of the farmers' time that might have been spent farming must be used to guard the crops against thieves. Similarly, some capital that might have been used for tractors is used for fences to keep out the thieves. The economy of Cornucopia gets much more output from its farmers and capital than does the economy of Kleptocopia. That is, Cornucopia has higher TFP.

This reasoning can help us rewrite the aggregate production function of an economy, like that used in Chapter 6 in equation (6.3), as

$$Y = IK^{\alpha}(hL)^{1-\alpha},$$

where I denotes the influence of an economy's social infrastructure on the productivity of its inputs. With this modification, we now have a complete theory of production that accounts for the empirical results documented in Chapter 3. Economies grow over time because new capital goods are invented and the agents in the economy learn to use the new kinds of capital (captured by h). However, two economies with the same K, h, and L may still produce different amounts of output because the economic environments in which those inputs are used to produce output differ. In one, capital may be used for fences, security systems, and pirate ships, and skills may be devoted to defrauding investors or collecting bribes. In another, all inputs may be devoted to productive activities.

7.6 THE CHOICE OF SOCIAL INFRASTRUCTURE

Why is the social infrastructure in some economies so much better than in others? Our questions about the determinants of long-run economic success are starting to resemble the beautiful *matrioshka* dolls of Russia,

in which each figurine contains another inside it. Each of our answers to the question of what determines long-run economic success seems to raise another question.

The questions also become increasingly difficult, and economists do not yet have firm answers about the determinants of social infrastructure. In the history of economic thought, the answers range far and wide. Max Weber argued in *The Protestant Ethic and the Spirit of Capitalism* (1976 [1920]) that belief systems were important and emphasized Protestantism's teachings regarding the individual. Other answers that have been proposed include culture or even climate and geography.

The question of what determines social infrastructure is one that has greatly concerned economic historian and 1993 Nobel Prize winner Douglass North in much of his research. A principle that has served North well is that individuals in power will pursue actions that maximize their own utility. Far from leaders being "benevolent social planners" who seek to maximize the welfare of the individuals in society, government officials are self-interested, utility-maximizing agents just like the rest of us. In order to understand why certain laws, rules, and institutions are put in place in an economy, we need to understand what the governors and the governees have to gain and lose and how easy it is for the governees to replace the governor. Applying this reasoning to the broad sweep of economic history, North (1981) states,

> From the redistributive societies of ancient Egyptian dynasties through the slavery system of the Greek and Roman world to the medieval manor, there was a persistent tension between the ownership structure which maximized the rents to the ruler (and his group) and an efficient system that reduced transaction costs and encouraged economic growth. This fundamental dichotomy is the root cause of the failure of societies to experience sustained economic growth [p. 25].

This same argument can help us to understand what Joel Mokyr (1990, p. 209) calls the "greatest enigma in the history of technology": why China was unable to sustain its technological lead after the fourteenth century. For several hundred years during the Middle Ages and culminating in the fourteenth century, China was the most technologically advanced society in the world. Paper, the shoulder-collar for harnessing horses, moveable type for printing, the compass, the clock, gunpowder, ship-building, the spinning wheel, and iron casting were

all invented in China centuries before they became known in the West. Yet by the sixteenth century many of these inventions had been either forgotten completely or simply left unimproved. It was the countries of western Europe rather than China that settled the New World and initiated the Industrial Revolution. Why? Historians disagree about the complete explanation, but a key factor is likely the lack of institutions supporting entrepreneurship.

What changed around the fourteenth century and led to the suppression of innovation and the demise of China's technological lead? One answer is the dynasty ruling China: the Ming dynasty replaced the Mongol dynasty in 1368. Mokyr, summarizing a plausible explanation advanced by several economic historians, writes,

> China was and remained an empire, under tight bureaucratic control. European-style wars between internal political units became rare in China after 960 A.D. The absence of political competition did not mean that technological progress could not take place, but it did mean that one decision maker could deal it a mortal blow. Interested and enlightened emperors encouraged technological progress, but the reactionary rulers of the later Ming period clearly preferred a stable and controllable environment. Innovators and purveyors of foreign ideas were regarded as troublemakers and were suppressed. Such rulers existed in Europe as well, but because no one controlled the entire continent, they did no more than switch the center of economic gravity from one area to another (Mokyr 1990, p. 231).

7.7 GROWTH MIRACLES AND DISASTERS

The government policies and institutions that make up the social infrastructure of an economy determine investment and productivity, and therefore also determine the wealth of nations. Fundamental changes in social infrastructure can then generate growth miracles and growth disasters.

Two classic examples are Japan and Argentina. From 1870 until World War II, Japan's income remained around 25 percent of U.S. income. After the substantial reforms put in place at the end of the war, Japanese relative income rose sharply, far beyond recovery back to the 25 percent level. Today, as a result of this growth miracle, Japanese income is roughly two-thirds that of income in the United States. Ar-

gentina is a famous example of the reverse movement — a growth disaster. Argentina was as rich as most western European countries at the end of the nineteenth century, but by 1997 income per worker had fallen to only 45 percent of that of the United States. Much of this decline is attributable to disastrous policy "reforms," including those of the Juan Perón era.

Why do such fundamental changes in social infrastructure occur? The answer probably lies in political economy and economic history. To predict when and whether such a change will occur in a particular economy surely requires detailed knowledge of the economy's circumstances and history. We can make progress by asking a slightly different question, however. Instead of considering the prospects for any individual economy, we can analyze the prospects for the world as a whole. Predicting the frequency with which such changes are likely to occur somewhere in the world is easier: we observe a large number of countries for several decades and can simply count the number of growth miracles and growth disasters.

A more formal way of conducting this exercise is presented in Table 7.1.[7] First, we sort countries into categories (or "bins") based on their 1960 level of GDP per worker relative to the world's leading economy (the United States during recent decades). For example, the bins correspond to countries with incomes of less than 5 percent of the world's leading economy, less than 10 percent but more than 5 percent, etc. Then, using annual data from 1960 to 1997 for 110 countries, we calculate the observed frequency with which countries move from one bin to another. Finally, using these sample probabilities, we compute an estimate of the long-run distribution of incomes.[8]

Table 7.1 shows the distribution of countries across the bins in 1960 and 1997, as well as an estimate of the long-run distribution. The results are intriguing. The basic changes from 1960 to 1997 have been documented in Chapter 3. There has been some "convergence" toward the United States at the top of the income distribution, and this phenomenon is evident in the table. The long-run distribution, according to the results shown in the table, strongly suggests that this convergence will play a dominant role in the continuing evolution of the income distribution. For example, in 1960 only 4 percent of countries had more than 80 percent of U.S. income and only 20 percent had more than 40 percent of U.S. income. In the long run, according to the results, 22 percent of countries will have relative incomes of more than 80 percent of the world's leading economy and 59 percent will have relative incomes of more than 40 percent. Similar changes are seen at the bottom of the distribution: in 1997, 24 percent of countries had less than 5 percent of U.S. income; in the long run, only 9 percent of countries are predicted to be in this category.

Several comments on these results are worth considering. First, what is it in the data that delivers the result? The basic answer to this question is apparent in Figure 3.6 of Chapter 3. Looking back at this figure, one sees that there are more countries moving up in the distribution than

[7] This section is drawn from Jones (1997). Quah (1993) first used this "Markov transition" approach to analyze the world income distribution.

[8] The sense in which this computation is different from that in Chapter 3 is worth emphasizing. There, we computed the steady state toward which each economy seems to be headed and examined the distribution of the steady states. Here, the exercise focuses much more on the very long run. In particular, according to the methods used to compute the long-run distribution in Table 7.1, if we wait long enough, there a is positive probability of any country ending up in any bin. This is discussed further in the coming examples.

TABLE 7.1 THE VERY-LONG-RUN DISTRIBUTION OF WORLD INCOME

"Bin"	Distribution 1960	Distribution 1997	Distribution Long-run	Years to "shuffle"
$\tilde{y} \leq .05$	16	24	9	513
$.05 < \tilde{y} \leq .10$	21	11	7	487
$.10 < \tilde{y} \leq .20$	25	15	9	379
$.20 < \tilde{y} \leq .40$	17	17	16	133
$.40 < \tilde{y} \leq .80$	16	24	37	330
$\tilde{y} > .80$	4	9	22	357

SOURCE: Calculations extending Jones (1997).

Note: Entries under "Distribution" reflect the percentage of countries with relative incomes in each "bin." "Years to shuffle" indicates the number of years after which the best guess as to a country's location is given by the long-run distribution, provided that the country begins in a particular bin.

moving down; there are more Italys than Venezuelas. In the last forty years, we have seen more growth miracles than growth disasters.

Second, the world income distribution has been evolving for centuries. Why doesn't the long-run distribution look roughly like the current distribution? This is a very broad and important question. The fact that the data say that the long-run distribution is different from the current distribution indicates that something in the world continues to evolve: the frequency of growth miracles in the last forty years must have been higher than in the past, and there must have been fewer growth disasters.

One possible explanation of this result is that society is gradually discovering the kinds of institutions and policies that are conducive to successful economic performance, and these discoveries are gradually diffusing around the world. To take one example, Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations* was not published until 1776. The continued evolution of the world income distribution could reflect the slow diffusion of capitalism during the last two hundred years. Consistent with this reasoning, the world's experiments with communism seem to be coming to an end only in the 1990s. Perhaps it is the diffusion of wealth-promoting institutions and social infrastructure that accounts for the continued evolution of the world income distribution. Moreover, there is no reason to think that the institutions in place today are the best possible institutions. Institutions themselves are simply "ideas," and it is very likely that better ideas are out there waiting to be found. Over the broad course of history, better institutions have been discovered and gradually implemented. The continuation of this process at the rates observed during the last forty years would lead to large improvements in the world income distribution.

The last column of Table 7.1 provides some insight regarding the length of time required to reach the long-run distribution. Consider shuffling a deck of playing cards right out of the pack—i.e., when they are initially sorted by suit and rank. How many shuffles does it take before the Ace of Spades has an equal probability of appearing anywhere in the deck? The answer turns out to be seven, provided the shuffles are perfect. Now suppose we consider a country in the richest income bin. How many years do we have to wait before the probability that the country is in a particular bin matches the probability implied by the long-run distribution? The last column of Table 7.1 reports that this number is 357 years. For a country starting from the poorest bin, it takes 513 years for initial conditions to cease to matter. These numbers are large, reflecting the fact that countries typically move very slowly through the world income distribution.

Other related experiments are informative. For example, one can calculate the frequency of "growth disasters." Although China was one of the most advanced countries in the world around the fourteenth century, today it has a GDP per worker of less than 10 percent that of the United States. What is the likelihood of such a dramatic change? Taking a country in the richest bin, only after more than 180 years is there a 10 percent probability that the country will fall to a relative income of less than 10 percent.

What about growth miracles? The "Korean experience" is not all that unlikely. A country in the 10 percent bin will move to an income level in the 40 percent bin or higher with a 10 percent probability after 46 years. The same is true of the "Japanese experience": a country in the 20 percent bin will move to the richest category with a 10 percent probability after 60 years. Given that there are a large number of countries in these initial categories, one would expect to see several growth miracles at any point in time.

7.8 SUMMARY

The *social infrastructure* of an economy — the rules and regulations and the institutions that enforce them — is a primary determinant of the extent to which individuals are willing to make the long-term investments in capital, skills, and technology that are associated with long-run economic success. Economies in which the government provides an environment that encourages production are extremely dynamic and successful. Those in which the government abuses its authority to engage in and permit diversion are correspondingly less successful.

Implicit in this theory of long-run economic performance is a theory that addresses the third fundamental question of economic growth discussed in the introduction of this book, the question of "growth miracles." How is it that some countries such as Singapore, Hong Kong, and Japan can move from being relatively poor to being relatively rich

over a span of time as short as forty years? Similarly, how is it that an economy like Argentina's or Venezuela's can make the reverse move?

This theory suggests that the answer is to be found in basic changes in the social infrastructure of the economy: changes in the government policies and institutions that make up the economic environment of these economies.

Why do some economies develop social infrastructures that are extremely supportive of production while others do not? Why was the Magna Carta written in England and why were its principles embraced throughout Europe? How did England develop a separation of powers between the Crown and Parliament and a strong judicial system? Why did the United States benefit from the Constitution and the Bill of Rights? And most important, why, given historical experience, have some economies successfully adopted these institutions and the social infrastructure associated with them while others have not? Fundamentally, these are the questions that must be addressed to understand the world pattern of economic success and how it changes over time.

EXERCISES

1. *Cost-benefit analysis.* Suppose an investment project yields a profit of \$100 every year, starting one year after the investment takes place. Assume the interest rate for computing present values is 5 percent.
   (a) If $F = \$1{,}000$, is the investment worth undertaking?
   (b) What if $F = \$5{,}000$?
   (c) What is the cutoff value for F that just makes the investment worthwhile?
2. *Can differences in the utilization of factors of production explain differences in TFP?* Consider a production function of the form $Y = IK^{\alpha}(hL)^{1-\alpha}$, where I denotes total factor productivity and the other notation is standard. Suppose I varies by a factor of 10 across countries, and assume $\alpha = 1/3$.
   (a) Suppose differences in infrastructure across countries lead only to differences in the fraction of physical capital that is utilized in production (versus its use, say, as fences to protect against diversion). How much variation in the utilization of capital do we need in order to explain the variation in TFP?
   (b) Suppose both physical capital and skills vary because of utilization, and for simplicity suppose that they vary by the same factor. How much variation do we need now?
   (c) What do these calculations suggest about the ability of utilization by itself to explain differences in TFP? What else could be going on?
3. *Social infrastructure and the investment rate.* Suppose that rates of return to capital are equalized across countries because the world is an open economy, and suppose that all countries are on their balanced growth paths. Assume the production function looks like $Y = IK^{\alpha}L^{1-\alpha}$, where I reflects differences in social infrastructure.
   (a) Show that differences in I across countries do not lead to differences in investment rates.
   (b) How might social infrastructure in general still explain differences in investment rates?
4. Discuss the meaning of the quotation that began this chapter.