November 7, 2004

Chapter 1

Thinking about Macroeconomics

In macroeconomics, we study the overall or aggregate performance of an economy. A lot of our focus will be on an economy's total output of goods and services, as measured by the **real gross domestic product** (**GDP**). We will study the breakdown of GDP into its major components: consumption, gross private domestic investment (purchases of new capital goods—equipment and structures—by the private sector), government consumption and investment, and net exports of goods and services. We also examine the aggregates of **employment** (persons with jobs) and **unemployment** (persons without jobs who are seeking work).

These terms refer to quantities of goods or labor. We are also interested in the prices that correspond to these quantities. For example, we consider the dollar prices of the goods and services produced in an economy. When we look at the price of the typical or average item, we refer to the **general price level**. We also study the **wage rate**, which is the dollar price of labor; the **rental price**, which is the dollar price paid to use capital goods; and the **interest rate**, which determines the cost of borrowing and the return to lending. When we consider more than one economy, we can study the exchange rate, which is the rate at which one form of money (say the euro) exchanges for another form of money (say the U.S. dollar).

We will set up an economic model, which allows us to study how the various quantities and prices are determined. We can use the model to see how the quantities and prices change in response to disturbances, such as a technological advance. We can also investigate the effects of the government's policies. For example, we consider monetary policy, which involves the determination of the quantity of money and the setting of interest rates. We also study fiscal policy, which pertains to the government's choices of expenditures, taxes, and fiscal deficits.

The performance of the overall economy matters for everyone because it influences incomes, job prospects, and prices. Thus, it is important for us—and even more important for government policymakers—to understand how the macroeconomy operates. Unfortunately, as is obvious from reading the newspapers, macroeconomics is not a settled scientific field. Although there is consensus on many issues—such as some of the determinants of long-run economic growth—there is also controversy about many topics—such as the sources of economic fluctuations and the short-run effects of monetary policy. The main objective of this book is to convey the macroeconomic knowledge that has been attained, as well as to point out areas in which a full understanding has yet to be achieved.

I. Output, Unemployment, and Prices in U.S. History

To get an overview of the subject, consider the historical record for some of the major macroeconomic variables in the United States. Figure 1.1 shows the total output of goods and services from 1869 to 2003. (The starting date is determined by the available data.) Our measure of aggregate output is the real gross domestic product (GDP).[1] This concept expresses quantities in terms of a base year, which happens to be 2000. Chapter 2 considers **national-income accounting** and thereby provides the conceptual details for measuring real GDP.

The general upward trend of real GDP in Figure 1.1 reflects the long-term growth of the U.S. economy. The average growth rate of real GDP from 1869 to 2003 was 3.5% per year. This growth rate meant that, over 134 years, the economy's total output of goods and services expanded 113-fold. If we divide through by population to determine real per capita GDP, we find that the average per capita growth rate was 2.0% per year. This rate equals the 3.5% per year growth rate of real GDP less the 1.5% per year growth rate of population. The growth rate of real per capita GDP of 2.0% per year meant that, over 134 years, the real GDP per capita expanded 15-fold.

Figure 1.2 shows the growth rate of real GDP for each year from 1870 to 2003. The year-to-year growth rates vary substantially around their mean of 3.5%. These variations are called **economic fluctuations** or, sometimes, the **business cycle**.[2] When real GDP falls toward a low point or trough, the economy is in a **recession** or an economic contraction. When real GDP expands toward a high point or peak, the economy is in a **boom** or an economic expansion.

The dates marked in Figure 1.2 correspond to the major U.S. recessions since 1869. There are many ways to classify periods of recession. In this graph we mark as years of recession the years of low economic growth. In chapter 6, we use a somewhat

[1] The graph uses a proportionate (logarithmic) scale, so that each unit on the vertical axis corresponds to the same percentage change in real GDP. Because of data availability, the numbers before 1929 refer to real gross national product (GNP). We discuss the relation between GDP and GNP in chapter 2.

[2] The term business cycle can be misleading because it suggests a more regular pattern of ups and downs in economic activity than actually appears in the data.

more sophisticated method to classify recessions. However, most of the classifications are the same as those shown in Figure 1.2.

Note in Figure 1.2 the **Great Depression** from 1930 to 1933, during which real GDP declined at 8% per year for four years. Other major recessions before World War II occurred in 1893-94, 1907-08, 1914, 1920-21, and 1937-38. In the post-World War II period, the main recessions were in 1958, 1974-75, and 1980-82.

For economic booms, note first in Figure 1.2 the high rates of economic growth during World Wars I and II and the Korean War. Peacetime periods of sustained high economic growth before World War II were 1875 to 1880, 1896 to 1906, much of the 1920s, and the recovery from the Great Depression from 1933 to 1940 (except for the 1937-38 recession). After World War II, the periods of sustained high economic growth were 1961 to 1973 (except for the brief recession in 1970), 1983 to 1989, and 1992 to 2000.

Another way to gauge recessions and booms is to consider the **unemployment rate**—the fraction of persons seeking work who have no job. Figure 1.3 shows the unemployment rate for each year from 1890 to 2003. Over this period, the median unemployment rate was 5.6%. (The mean was 6.3%.) During recessions, the unemployment rate typically rises above its median. The extreme is the Great Depression, during which the unemployment rate reached 22% in 1932. Also noteworthy in the pre-World War II period are the average unemployment rates of 18% for 1931-35, 12% for 1938-39, 11% for 1894-98, and 8% for 1921-22. In the post-World War II period, the highest unemployment rate was 10% in 1982-83. Other periods of high unemployment rates were 8% in 1975-76 and 7% in 1958, 1961, and 1991-93.

Figures 1.2 and 1.3 show the turbulence of the U.S. economy during the two world wars and the 1930s. But suppose that we abstract from these extreme episodes and compare the post-World War II period with the one before World War I. Then the major message from the data is the similarity between the post-World War II and pre-World War I periods.

The average growth rate of real GDP was 3.4% per year from 1948 to 2003, compared with 3.8% from 1870 to 1914 and 3.4% from 1890 to 1914. The median unemployment rates were 5.6% from 1948 to 2003 and 5.2% from 1890 to 1914. (The means were 5.6% and 6.4%, respectively.) The extent of economic fluctuations—in the sense of the variability of growth rates of real GDP or of unemployment rates—was only moderately larger in the pre-World War I period than in the post-World War II period.[3] The economy has, of course, changed greatly over the 134 years from 1869 to 2003—including a larger role for government, a diminished share of agriculture in the GDP, and dramatic changes in the monetary system. Nevertheless, the U.S. data do not reveal major changes in the intensity of economic fluctuations or in the average rate of economic growth.

Figure 1.4 shows the evolution of the U.S. price level from 1869 to 2003. This graph measures the price level as the deflator for the GDP—we discuss the details of this price index in chapter 2. One striking observation is the persistent rise in the price level since World War II, contrasted with the movements up and down before World War II. There are long periods in the earlier history—1869 to 1892 and 1920 to 1933—during which the price level fell persistently.

---

[3] For a detailed comparison of real GDP and unemployment rates for the two periods, see Christina Romer (1986, 1987, 1988).

Figure 1.5 shows the annual **inflation rate** from 1870 to 2003. Each year's inflation rate is calculated as the growth rate of the price level shown in Figure 1.4. Notice that the inflation rates since World War II were all greater than zero, except for 1949. In contrast, many of the inflation rates before World War II were less than zero. Note also in the post-World War II period that the inflation rate fell sharply from a peak of 8.8% in 1980-81 to an average of 2.5% from 1983 to 2003 and only 1.9% from 1992 to 2003.

In subsequent chapters, we shall relate the changing behavior of the inflation rate to the changing character of monetary institutions and monetary policy. In the pre-World War II period, a key element is the **gold standard**—which the United States adhered to from 1879 until World War I and, to some extent, from World War I until 1933. In the post-World War II period, a key element is the changing monetary policy of the U.S. Federal Reserve. Notably, since the mid 1980s, the Federal Reserve and other major central banks have successfully pursued a policy of low and stable inflation.

II. Economic Models

As mentioned, we want to understand the determinants of major macroeconomic variables, such as real GDP and the general price level. To carry out this mission, we will construct a macroeconomic model. A model can be a group of equations or graphs or a set of conceptual ideas. We will use all of these tools in this book—some equations but, more often, graphs and ideas.

An economic model deals with two kinds of variables: endogenous variables and exogenous variables. The **endogenous variables** are the ones that we want the model to explain. For example, the endogenous variables in our macroeconomic model include real GDP, investment, employment, the general price level, the wage rate, and the interest rate.

The **exogenous variables** are the ones that a model takes as given and does not attempt to explain. A simple example of an exogenous variable is the weather (at least in models that do not allow for feedback from the economy to climate change). In many cases, the available technologies will be exogenous. For a single economy, the exogenous variables include the world prices of commodities such as oil and wheat, as well as levels of income in the rest of the world. In many cases, we will treat government policies as exogenous variables—for example, choices about monetary policy and about the government's spending and taxes. In particular, we will treat conditions of war and peace—which have important macroeconomic consequences—as exogenous variables.

The central idea of a model is that it tells us how to go from the exogenous variables to the endogenous variables. Figure 1.6 illustrates this process. We take as given the group of exogenous variables shown in the left box in the diagram. The model tells us how to go from these exogenous variables to the group of endogenous variables, shown in the right box in the diagram. Therefore, we can use the model to predict how changes in the exogenous variables affect the endogenous variables.

In macroeconomics, we are interested in the determination of macroeconomic—that is, economy-wide aggregate—variables, such as real GDP. However, to construct a macroeconomic model, we shall find it useful to build on a microeconomic approach to the actions of individual households and businesses. This microeconomics investigates individual decisions about how much to consume, how much to work, and so on. Then we can add up or aggregate the choices of individuals to construct a macroeconomic model. The underlying microeconomic analysis is called **microeconomic foundations**. These foundations will help us build a useful macroeconomic model.

A. A simple example—the coffee market

To illustrate general ideas about models and markets, we can start with the market for a single product, coffee. Our analysis will focus on three key tools used by economists: demand curves, supply curves, and market-clearing conditions (demand equals supply).

Individuals decide how much coffee to buy, that is, the quantity of coffee to demand. Influences on this demand include the individual's income, the price of coffee, P_c, and the price of a substitute good, say P_T, the price of tea. Since each individual is a negligible part of the coffee and tea markets, it makes sense that each individual would neglect the effect of his or her coffee and tea consumption on P_c and P_T. That is, each individual is a **price-taker**; he or she just decides how much coffee and tea to buy at given prices, P_c and P_T. This assumption about price-taking behavior is called **perfect competition**.

Reasonable behavior for an individual suggests that the individual's quantity of coffee demanded would rise with income, fall with the coffee price, P_c, and rise with the price of the substitute good, P_T. These results for individuals are examples of microeconomic analysis. If we add up across all individuals, we can determine the aggregate quantity of coffee demanded as a function of aggregate income, denoted by Y, and the prices P_c and P_T. We can isolate the effect of the coffee price, P_c, on the total quantity of coffee demanded by drawing a market **demand curve.** This curve shows the total quantity of coffee demanded, Q_c^d, as a function of P_c.

Figure 1.7 shows the market demand curve for coffee. As already noted, a decrease in P_c increases Q_c^d. Recall, however, that the demand curve applies for given values of aggregate income, Y, and the price of tea, P_T. If Y rises, the quantity of coffee demanded, Q_c^d, increases for a given price, P_c. Therefore, the demand curve shown in Figure 1.7 would shift to the right. If P_T falls, the quantity of coffee demanded, Q_c^d, decreases for a given price, P_c. Therefore, the demand curve would shift to the left.

We also have to consider how individual producers of coffee decide how much to offer for sale on the market, that is, how much coffee to supply. Influences on this supply include the price of coffee, P_c, and the cost of producing additional coffee. We assume, as in our analysis of demand, that the suppliers of coffee are price-takers with respect to P_c. This assumption could be questioned because some producers of coffee are large and might consider the effects of their actions on P_c. However, an extension to allow for this effect would not change our basic analysis of the market for coffee.

Reasonable behavior by an individual producer suggests that the quantity of coffee supplied would rise with the price of coffee, P_c, and fall with an increase in the cost of producing additional coffee. For example, bad weather that destroys part of the coffee crop in Brazil would raise the cost of producing coffee and, thereby, reduce the coffee supplied by Brazilians. These results for individual producers are examples of microeconomic analysis.

If we add up across all producers, we can determine the aggregate quantity of coffee supplied. One result is that a rise in P_c increases the aggregate quantity of coffee supplied, Q_c^s. The total quantity supplied also depends on weather conditions in coffee-producing areas, such as Brazil and Colombia.

As in our analysis of demand, we can isolate the effect of the coffee price, P_c, on the total quantity of coffee supplied by drawing a market **supply curve**. This curve, shown in Figure 1.8, gives the total quantity of coffee supplied, Q_c^s, as a function of P_c. As already noted, an increase in P_c raises Q_c^s. This supply curve applies for given cost conditions for producing coffee, in particular, for given weather in coffee-producing areas. If bad weather destroys part of Brazil's coffee crop, the market quantity of coffee supplied, Q_c^s, decreases for a given price, P_c. Therefore, the supply curve shown in Figure 1.8 would shift to the left.

Demand and supply curves are functions

The market demand for coffee can be written as a function,

$$Q_c^d = D(P_c, Y, P_T).$$

The function $D(\cdot)$ determines the quantity of coffee demanded, Q_c^d, for any specified values of the three demand determinants, P_c, Y, and P_T. We assume that the function $D(\cdot)$ has the properties that Q_c^d decreases with the price of coffee, P_c, rises with income, Y, and rises with the price of tea, P_T. Figure 1.7 graphs Q_c^d against P_c for given values of the other demand determinants, Y and P_T. It is important to distinguish the demand curve, $D(\cdot)$, shown in Figure 1.7, from the quantity demanded, Q_c^d, at a given price, P_c (and for given Y

and P_T). The demand curve refers to the whole functional relationship between quantity demanded and price, *D(·)*, whereas the quantity demanded, Q_c^d, refers to one of the points along the curve.

The market supply of coffee is also a function, which can be written as

$$Q_c^s = S(P_c, \textit{weather}).$$

We assume that the function *S(·)* has the properties that the quantity supplied, Q_c^s, rises with P_c and with better weather in coffee-producing areas. Figure 1.8 graphs the quantity supplied, Q_c^s, against P_c, for given weather conditions. It is important to remember that the supply curve, *S(·)*, refers to the whole functional relationship between quantity supplied and price, whereas the quantity supplied, Q_c^s, refers to one of the points along the curve.

Figure 1.9 shows the clearing of the market for coffee. The price of coffee, P_c, is assumed to adjust to equate the quantity supplied, Q_c^s, to the quantity demanded, Q_c^d. This market-clearing price is the value $(P_c)^*$ shown in the figure. The corresponding market-clearing quantity of coffee is $(Q_c)^*$.

Why do we assume that the coffee price, P_c, adjusts to the market-clearing value, $(P_c)^*$? For any other price, the quantities supplied and demanded would be unequal. For example, if P_c were less than $(P_c)^*$, the quantity demanded, Q_c^d, would be greater than the quantity supplied, Q_c^s. In that case, some coffee drinkers must be unsatisfied; they would not be able to buy the quantity of coffee that they want at the price P_c. That is, suppliers would be unwilling to provide enough coffee to satisfy all of the desired purchases at this low price. In this circumstance, competition among the eager demanders of coffee would cause the market price, P_c, to rise toward $(P_c)^*$.

Conversely, if P_c were higher than $(P_c)^*$, the quantity demanded, Q_c^d, would be less than the quantity supplied, Q_c^s. In this case, some coffee producers must be unsatisfied; they would not be able to sell the full quantity of coffee that they want to sell at the price P_c. That is, coffee drinkers would be unwilling to buy all of the coffee that the producers offer at this high price. In this situation, competition among the eager suppliers of coffee would cause the market price, P_c, to fall toward $(P_c)^*$.

The market-clearing price, $P_c = (P_c)^*$, is special because only at this price is there no pressure for the coffee price to rise or fall. In this sense, the market-clearing price is an **equilibrium** price. This price tends to remain the same unless there are shifts to the demand curve or the supply curve.

We can think of our market-clearing analysis of the coffee market as a *model* of how the coffee market operates. The two endogenous variables in the model are the price, P_c, and quantity, Q_c, of coffee. We can use the market-clearing analysis from Figure 1.9 to see how changes in exogenous variables affect the endogenous variables in the model. The exogenous variables are the outside forces that shift the demand and supply curves for coffee. For demand, we referred to two exogenous variables: income,

Y, and the price of tea, P_T.[4] For supply, we mentioned as exogenous variables the weather conditions in coffee-producing areas, such as Brazil.

Figure 1.10 shows how an increase in demand affects the coffee market. The rise in demand could reflect an increase in income, Y, or the price of tea, P_T. We represent the increase in demand by a rightward shift of the demand curve. That is, consumers want to buy more coffee at any given price, P_c. We see from the diagram that the market-clearing price rises from $(P_c)^*$ to $(P_c)^{*\prime}$, and the market-clearing quantity increases from $(Q_c)^*$ to $(Q_c)^{*\prime}$. Thus, our model of the coffee market predicts that increases in Y or P_T raise P_c and Q_c. As in the conceptual diagram shown in Figure 1.6, the model tells us how changes in the exogenous variables affect the endogenous variables.

Figure 1.11 shows how a decrease in supply affects the coffee market. The reduction in supply could reflect poor weather conditions in coffee-producing areas, such as Brazil and Colombia. We represent the decrease in supply by a leftward shift of the supply curve. That is, producers want to sell less coffee at any given price, P_c. We see from the diagram that the market-clearing price rises from $(P_c)^*$ to $(P_c)^{*\prime}$, and the market-clearing quantity decreases from $(Q_c)^*$ to $(Q_c)^{*\prime}$. Thus, our model of the coffee market predicts that a poor coffee harvest raises P_c and lowers Q_c.

Table 1.1 summarizes the results from the market-clearing model of the coffee market. As in Figure 1.6, the model tells us how changes in the exogenous variables affect the endogenous variables.

[4] From a broader perspective that includes the tea market and the overall economy, the price of tea, P_T, and incomes would also be endogenous variables. This broader analysis is called general-equilibrium theory—that is, it considers the conditions for the clearing of all markets simultaneously. The limitation to a single market, such as the one for coffee, is an example of partial-equilibrium analysis. In this case, we assess the clearing of the coffee market while taking as given the outcomes in all the other markets.

Our macroeconomic model will use this kind of market-clearing analysis to predict how changes in exogenous variables affect the endogenous macroeconomic variables. However, we will not study an array of goods, such as coffee and other

Table 1.1

Effects of Changes in Exogenous Variables on the Endogenous Variables in the Coffee Market

Change in exogenous variable	Effect on P_c	Effect on Q_c
Increase in income, *Y*	rises	rises
Increase in price of tea, P_T	rises	rises
Poor coffee harvest	rises	falls

products. Rather, we will consider the aggregate demand for and supply of a composite good that corresponds to the economy's overall output, the real GDP. We will also analyze the demand for and supply of factors of production—labor and capital services.

B. Flexible versus sticky prices

When we studied the market for coffee, we focused on market-clearing conditions. Therefore, when an exogenous variable changed, we based our predictions on how this change altered the market-clearing price and quantity. The assumption that underlies this analysis is that the price of coffee adjusts rapidly to clear the market for coffee, that is, to equate the quantity demanded to the quantity supplied. We observed that, if the price differed from its market-clearing value, either demanders or suppliers of coffee could not be satisfied in their offers to buy or sell at the established price. Consequently, there was always pressure for the coffee price to adjust toward its market-clearing value—the market-clearing price was the only equilibrium price.

Although economists would generally accept the focus on market-clearing prices when analyzing coffee or similar products, there is less agreement on whether macroeconomics should focus on market-clearing conditions. In particular, not all economists agree that we should consider only situations of market clearing in the market for the composite good that represents real GDP or in the market for labor. For long-run analysis, there is a consensus that a market-clearing framework provides the best guide to how an economy evolves. Therefore, in our study of long-run economic growth in chapters 3-5, we use a market-clearing, equilibrium approach. However, for analyses of short-run macroeconomic fluctuations, there is a sharp divide among economists as to whether a market-clearing model provides useful insights.

The famous economist John Maynard Keynes, writing in the wake of the Great Depression in the 1930s, argued that the labor market typically did not clear—he thought that the labor market was usually in a state of **disequilibrium**. In particular, he argued that wage rates were sticky and adjusted only slowly to generate equality between the quantities of labor demanded and supplied. More recently, some macroeconomists have emphasized instead the tendency of some goods markets to be in disequilibrium. This approach, called **New Keynesian Economics**, argues that some prices are sticky and move only slowly to equate the quantities of goods demanded and supplied.

Other economists argue that an equilibrium approach, which relies on market-clearing conditions, can explain a lot about short-run economic fluctuations. The most well-known application of this idea is called **real business cycle theory**. This approach applies essentially the same methodology to understanding short-run fluctuations that most economists apply to the analysis of long-run economic growth. Wages and prices are viewed as sufficiently flexible in the short run so that a useful macroeconomic analysis can concentrate on market-clearing positions. As in our analysis of the coffee market (summarized in Table 1.1), we can then focus on how changes in exogenous variables affect market-clearing quantities and prices.

One point that seems clear is that we cannot understand or evaluate sticky-price models unless we have the flexible-price, market-clearing model as a benchmark. After all, macroeconomists agree that the economy is always approaching the market-clearing solution—that is why this setting is the one typically used to study long-run economic growth. A reasonable inference is that, whatever the ultimate verdict on the significance of sticky prices in the short run, it is best to begin macroeconomic analysis with a market-clearing model.

We set out the basic market-clearing framework in chapters 6-10. Then we extend this setting to allow for inflation in chapter 11 and for government spending, taxes, and fiscal deficits in chapters 12-14. Chapter 15 allows for misperceptions about prices and wages but continues to assume a market-clearing framework. Only in chapter 16 are we ready to assess the sticky wages and prices that are the hallmarks of Keynesian and New Keynesian models.

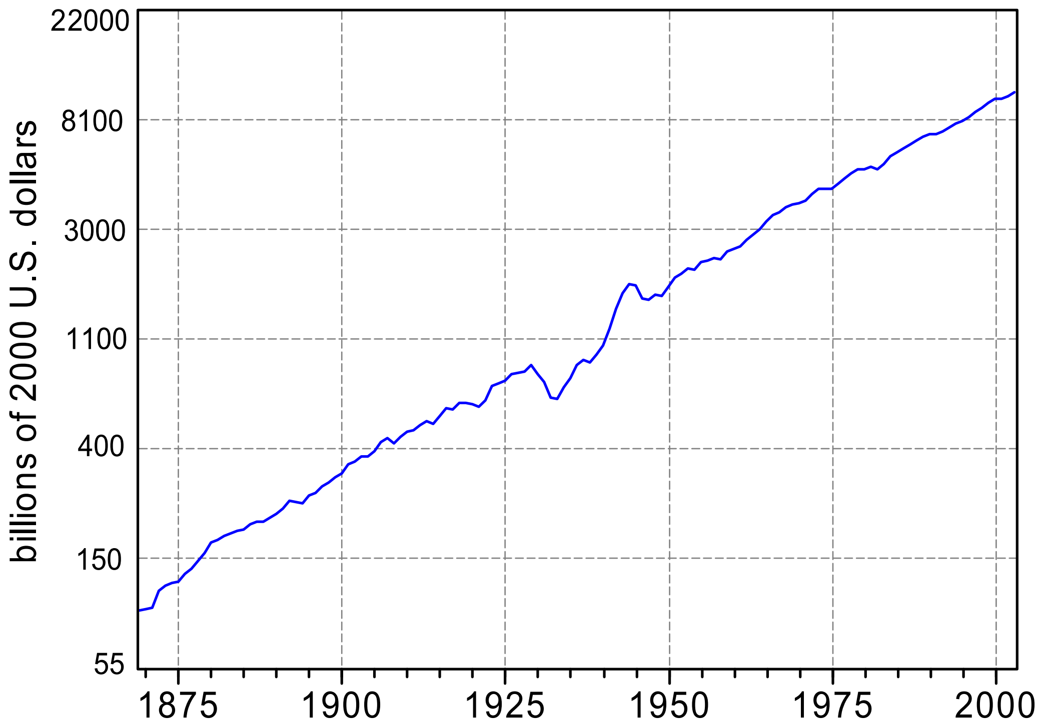


Figure 1.1

U.S. Real GDP, 1869-2003

The graph shows the real gross domestic product (GDP) on a proportionate (logarithmic) scale. Data before 1929 are for real gross national product (GNP). The numbers are in billions of 2000 U.S. dollars.

Sources: Data since 1929 are from Bureau of Economic Analysis. Values from 1869 to 1928 are based on data in Christina Romer (1987, 1988).

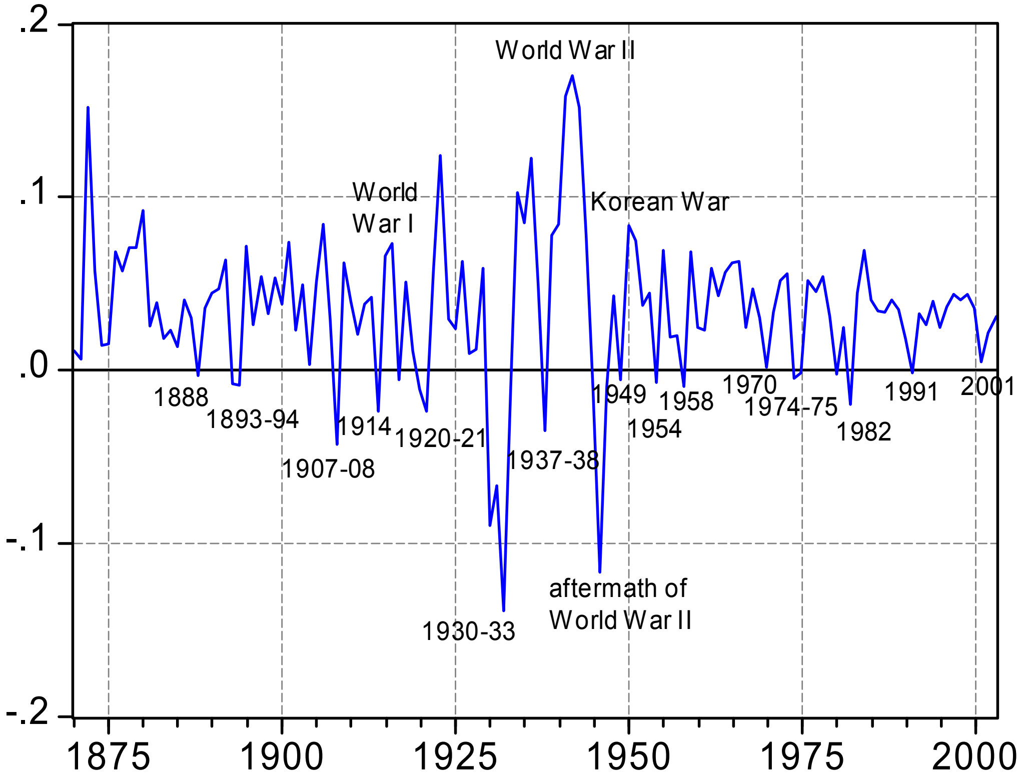


Figure 1.2

Growth Rate of U.S. Real GDP, 1870-2003

The graph shows the annual growth rate of real GDP (real GNP before 1929). The growth rates are calculated from the values of real GDP (or real GNP) shown in Figure 1.1. Aside from the years of major war, the years marked are recession periods. These periods have low (typically negative) rates of economic growth.

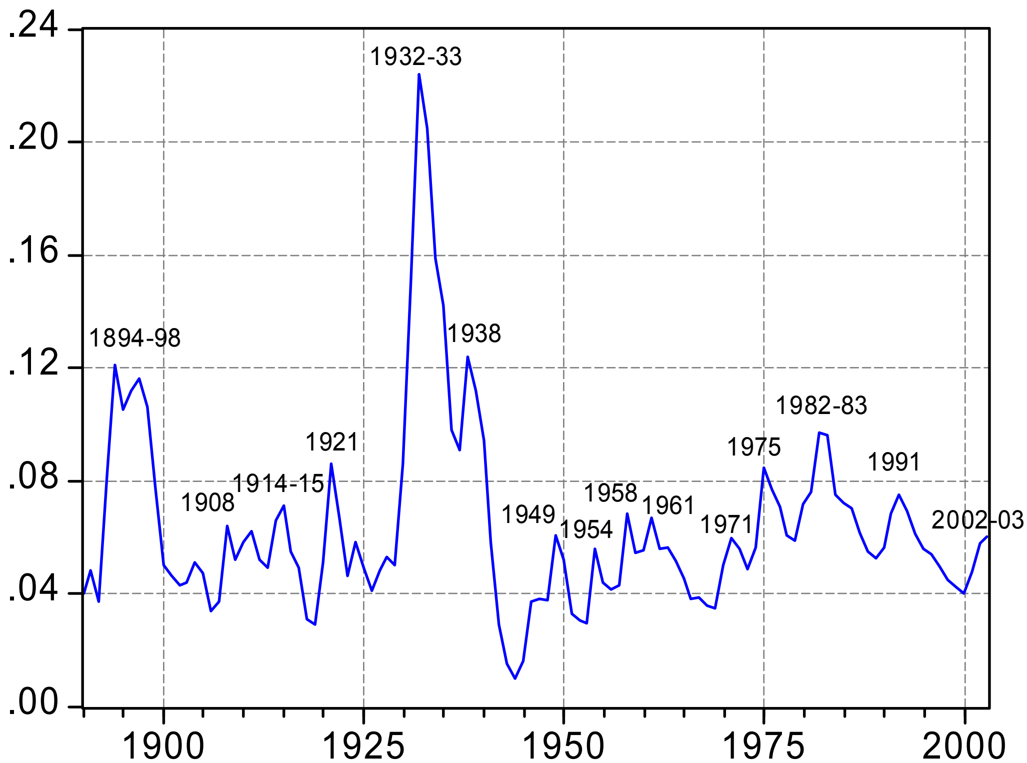


Figure 1.3

U.S. Unemployment Rate, 1890-2003

The graph shows the U.S. unemployment rate.

Sources: Data since 1929 are from Bureau of Labor Statistics. Values from 1890 to 1928 are based on data in Christina Romer (1986, Table 9). Values for 1933-43 were adjusted to classify federal emergency workers as employed, as discussed in Michael Darby (1976).

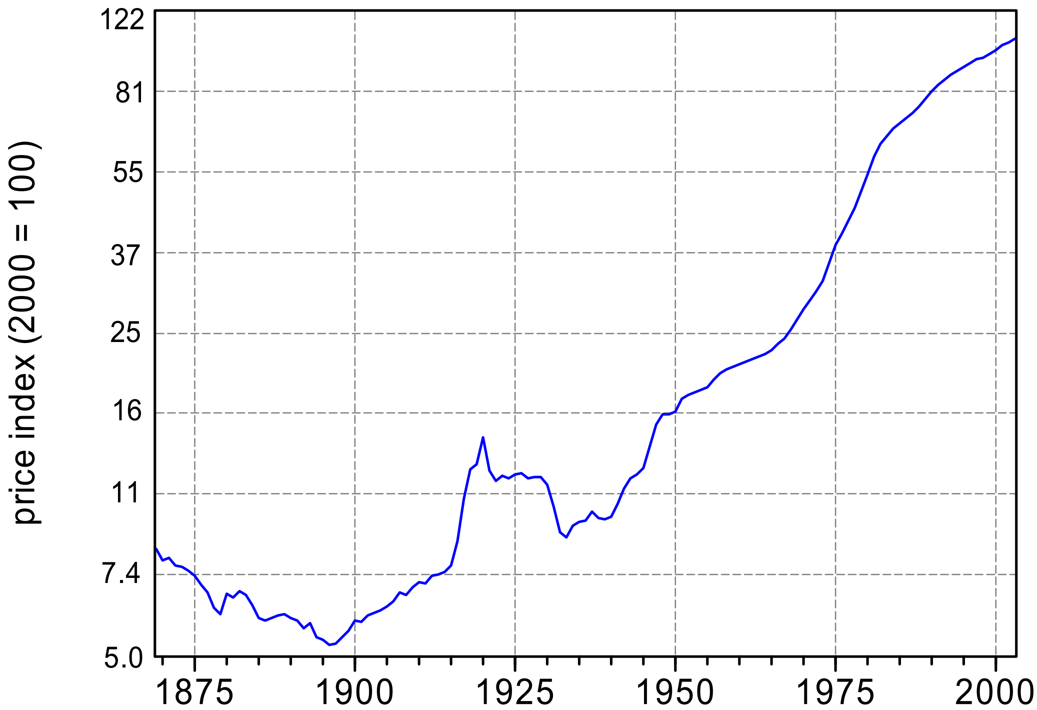


Figure 1.4

U.S. Price Level, 1869-2003

The graph shows the price deflator for the GDP (GNP before 1929). The numbers are on a proportionate (logarithmic) scale, with the value for the year 2000 set at 100. The sources are those indicated for GDP in Figure 1.1.

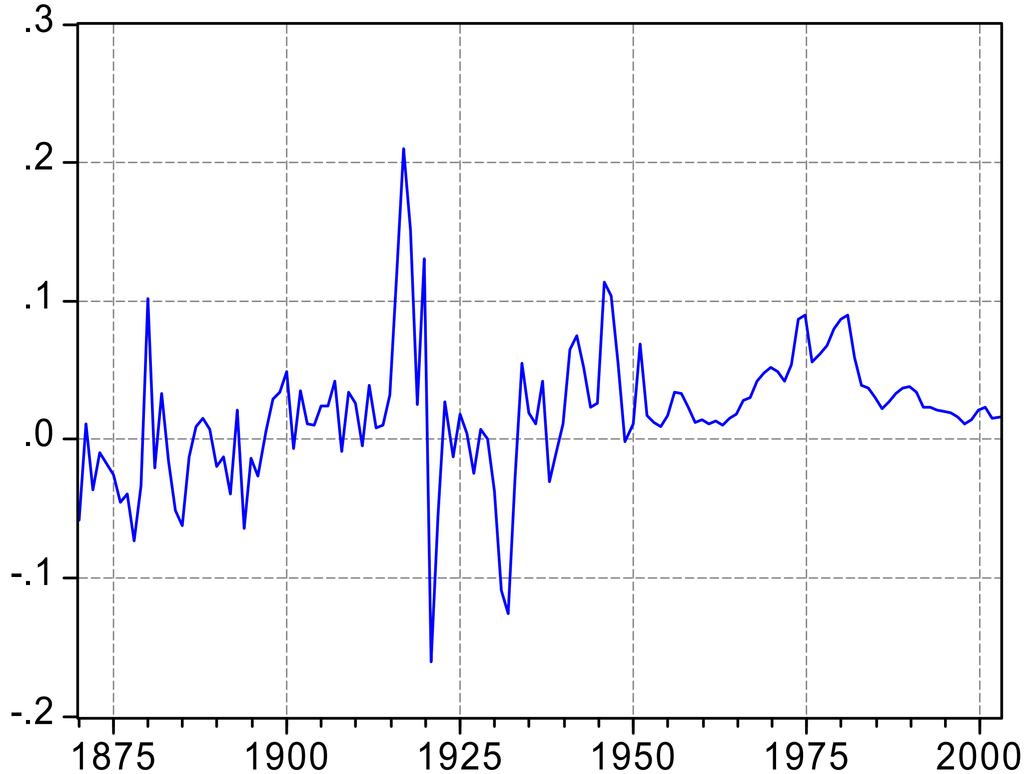


Figure 1.5

U.S. Inflation Rate, 1870-2003

The graph shows the annual inflation rate based on the GDP deflator (GNP deflator before 1929). The inflation rate is calculated as the annual growth rate of the price level shown in Figure 1.4.

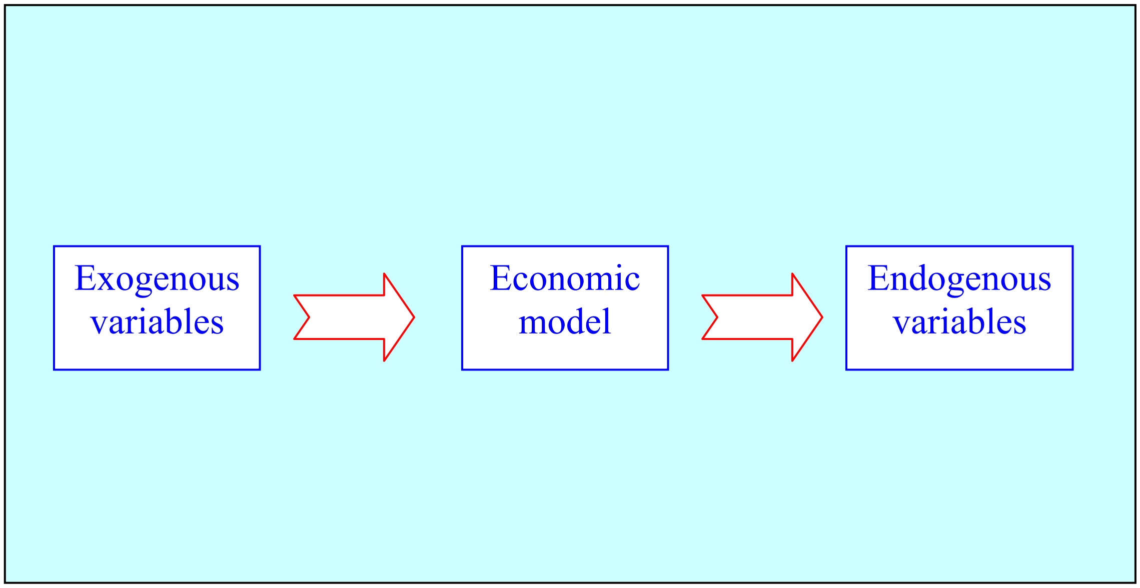


Figure 1.6

The Workings of an Economic Model

A model is a theory that tells us how to go from a group of exogenous variables to a group of endogenous variables. The model may be a list of equations or graphs or a set of conceptual ideas. The exogenous variables come from outside the model and are therefore not explained by the model. The endogenous variables are the ones that the model seeks to explain. With the help of the model, we can predict how changes in the exogenous variables affect the endogenous variables.

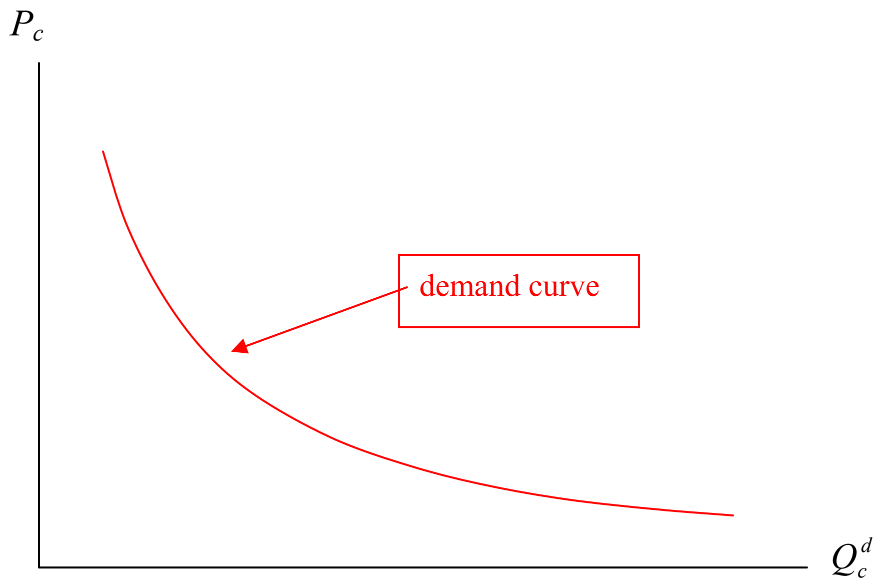


Figure 1.7

Demand Curve for Coffee

The market demand curve shows the total quantity of coffee demanded, Q_c^d, as a function of the price of coffee, P_c. A decrease in P_c raises Q_c^d. The demand curve applies for given aggregate income, Y, and the price of tea, P_T. If Y rises, the quantity of coffee demanded, Q_c^d, increases for given P_c. Therefore, the demand curve in the diagram would shift to the right. If P_T falls, the quantity of coffee demanded, Q_c^d, decreases for given P_c. Therefore, the demand curve shifts to the left.

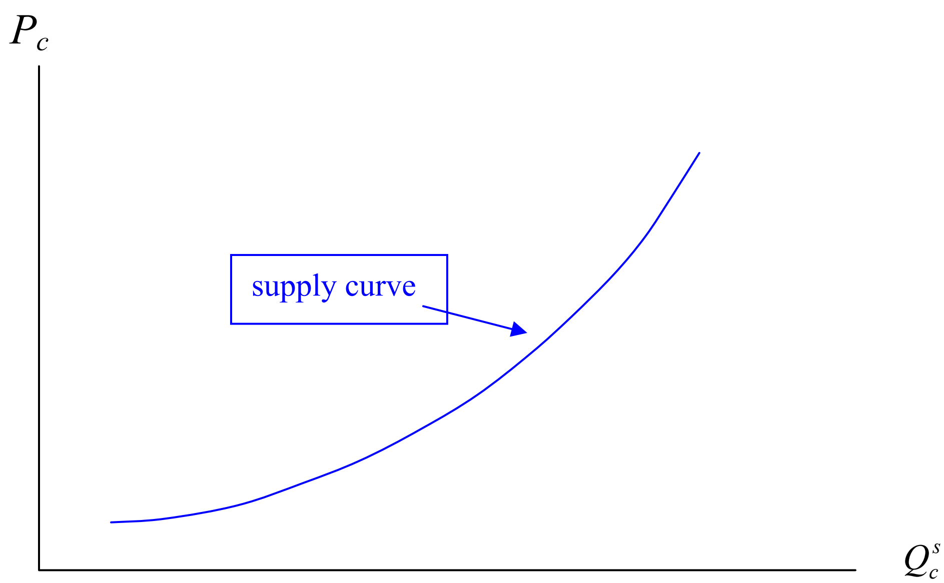


Figure 1.8

Supply Curve for Coffee

The market supply curve shows the total quantity of coffee supplied, Q_c^s, as a function of the price of coffee, P_c. A increase in P_c raises Q_c^s. The supply curve applies for given conditions that affect the cost of producing additional coffee. For example, a harvest failure in Brazil would decrease the total quantity of coffee supplied, Q_c^s, for a given price, P_c. Therefore, the supply curve would shift to the left.

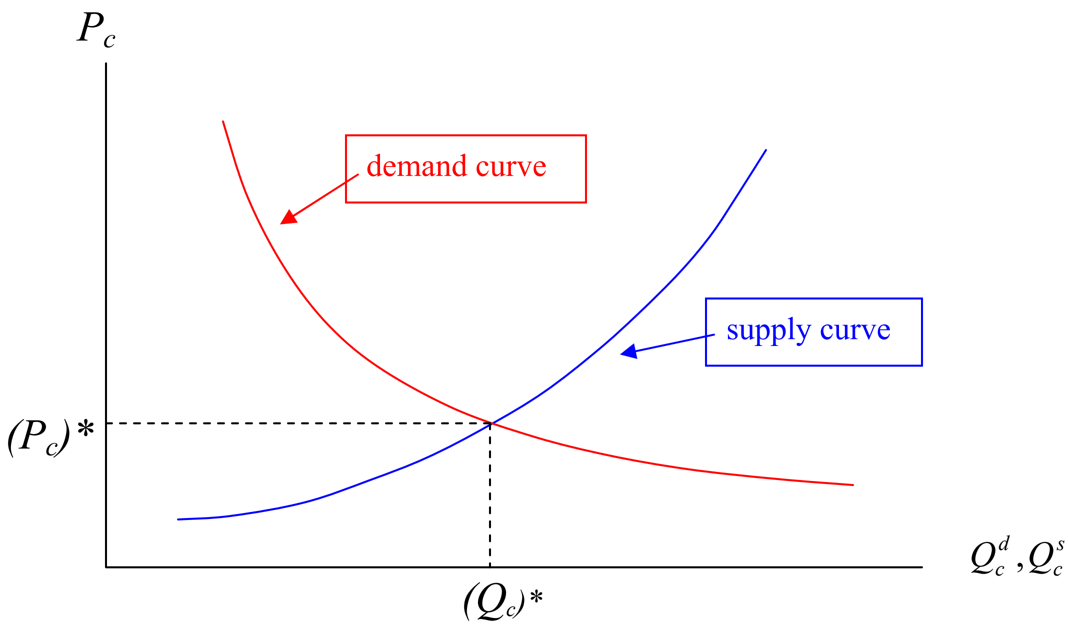


Figure 1.9

Clearing of the Market for Coffee

The coffee market clears at the price $(P_c)^*$ and quantity $(Q_c)^*$. At this point, the quantity of coffee supplied equals the quantity demanded.

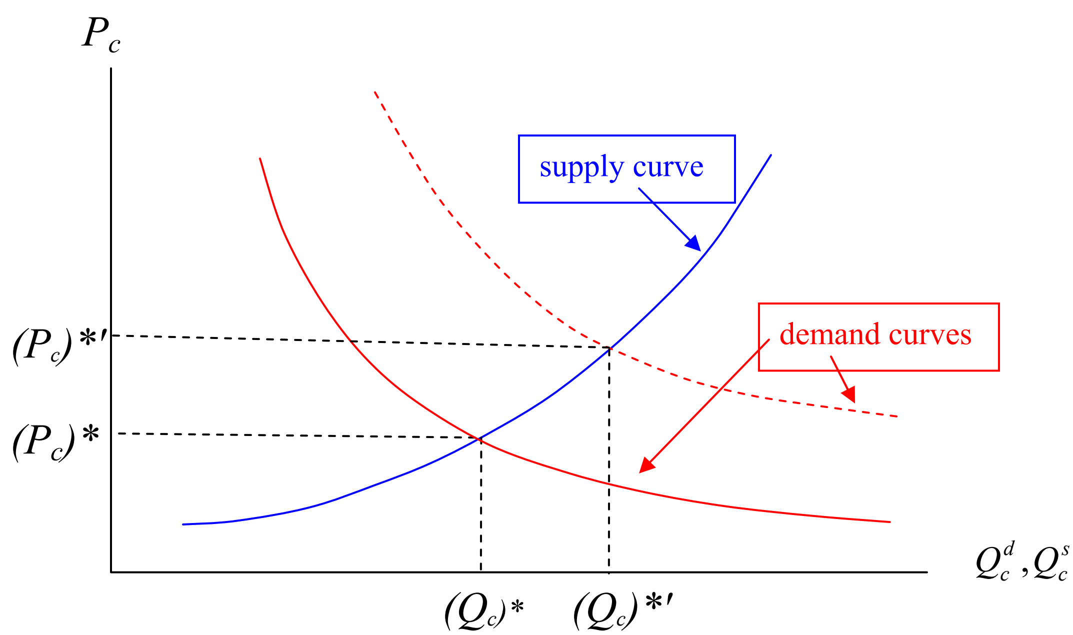


Figure 1.10

Effect of an Increase in Demand on the Coffee Market

In Figure 1.9, the coffee market cleared at the price $(P_c)^*$ and quantity $(Q_c)^*$. An increase in income, Y, or in the price of tea, P_T, raises the demand for coffee. Therefore, the demand curve shifts rightward from the solid red curve to the dashed red curve. The market-clearing price of coffee rises to $(P_c)^{*\prime}$, and the market-clearing quantity of coffee rises to $(Q_c)^{*\prime}$.

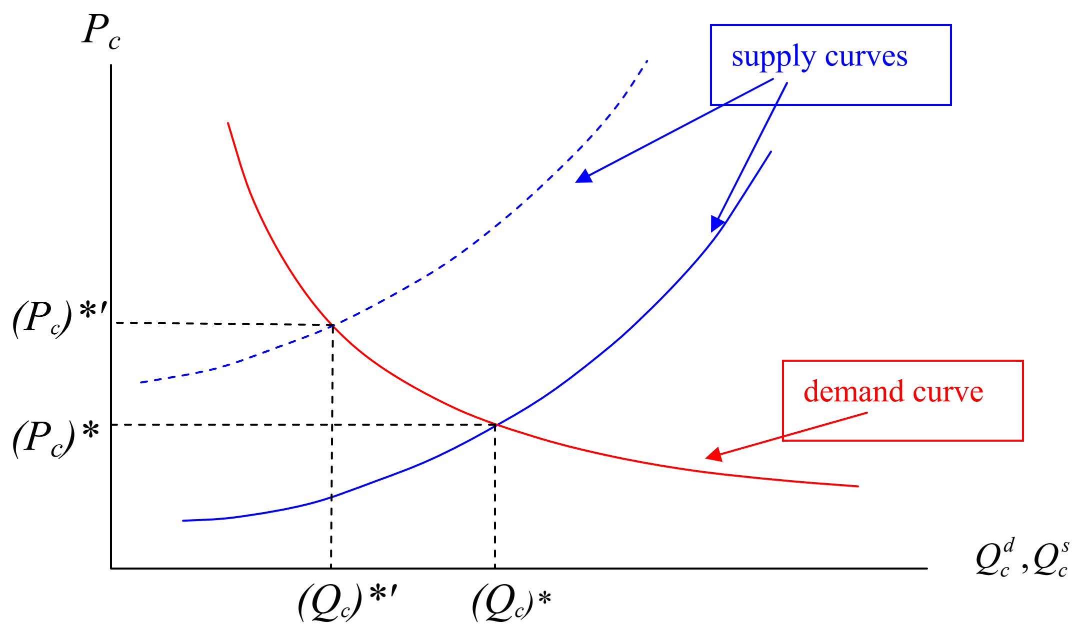


Figure 1.11

Effect of a Decrease in Supply on the Coffee Market

In Figure 1.9, the coffee market cleared at the price $(P_c)^*$ and quantity $(Q_c)^*$. A harvest failure in Brazil reduces the supply of coffee. Therefore, the supply curve shifts leftward from the solid blue curve to the dashed blue curve. The market-clearing price of coffee rises to $(P_c)^{*\prime}$, and the market-clearing quantity of coffee falls to $(Q_c)^{*\prime}$.

October 17, 2004

Chapter 2

National-Income Accounting:
Gross Domestic Product and the Price Level

In chapter 1, we used terms such as gross domestic product (GDP) and the price level without defining them precisely. Now, by looking at national-income accounting, we develop the meanings of these terms. There are many challenging issues that arise in the construction of the national-income accounts. However, for our purposes, we have to deal only with the basic concepts.

I. Nominal and Real GDP

We begin with the gross domestic product or GDP. Nominal GDP measures the dollar (or euro, etc.) value of all the goods and services that an economy produces during a specified period, such as a year. For example, in 2003, the U.S. nominal GDP was 11.0 trillion U.S. dollars. The nominal GDP is a **flow** concept. It measures the dollar amount of goods produced per unit of time, for example, per year.

Consider the definition of nominal GDP one step at a time. The word nominal means that the goods produced during a year are measured in terms of their value in dollars (or in units of another currency, such as the euro or yen). For most goods and services—pencils, automobiles, haircuts, and so on—the dollar value is determined by the price at which these items sell in the marketplace.

Some goods and services, notably those produced by governments, are not exchanged on markets. For example, the government does not sell its services for national defense, the justice system, and police. These items enter into nominal GDP at their nominal (dollar) cost of production. This treatment is problematic because it amounts to assuming that government employees experience no changes over time in their productivity. However, in the absence of market prices, it is unclear what alternative approach would be more accurate.

Another important item, owner-occupied housing, enters into GDP in accordance with an estimate of what this housing would fetch on the market if the owner rented the property to another person. This amount is called the **imputed rental income** on owner-occupied housing. Conceptually, the same approach ought to apply to consumer durables, for example, to households' automobiles, furniture, and appliances. However, this treatment has not been followed, that is, the GDP does not include an estimated rental income on consumer durables.[1] For government-owned property, the assumption in the national accounts is that the imputed rental income equals the estimated depreciation of the property. This assumption is problematic but, again, a preferred alternative method is not obvious.

It is important to understand that the nominal GDP includes the value of the goods and services produced during a specified time interval, such as a year. That is, GDP measures current production. For example, if an auto-maker manufactures and sells a new car in 2005, the full value of the car counts in the GDP for 2005. However, if

[1] For capital owned by businesses, the rental value of the capital—say factories and machinery—contributes to the measured value of business output. Therefore, to measure GDP, it is unnecessary to estimate an imputed rental income on business capital. The market value of output is all that we need.

someone sells in 2005 a used car that was built in 2004, this sale does not count in the GDP for 2005.

The nominal GDP can be misleading because it depends on the overall level of prices, as well as on the physical quantity of output. Table 2.1 illustrates this problem. Think about a simple economy that produces only butter and golf balls. The table shows the hypothetical quantities and prices of these goods in 2005 and 2006. In 2005, the economy produces 50 pounds of butter, which sells at $2.00 per pound. Thus, the dollar value of 2005's butter output is $100. Similarly, the economy produces 400 golf balls, priced at $1.00 per ball, for a golf-ball output of $400. The nominal GDP for 2005 is therefore the sum of the dollar values of butter and golf-ball output: $100 + $400 = $500.

The columns labeled 2006a and 2006b show two possibilities for prices and outputs in 2006. In case *a*, the prices of both goods rise—to $3.00 per pound of butter and $1.10 per golf ball. In case *b*, the prices of both goods decline—to $1.50 per pound of butter and $0.89 per golf ball. In case *a*, the quantities of both goods decline—to 40 pounds of butter and 391 golf balls. In case *b*, the quantities of both goods rise—to 70 pounds of butter and 500 golf balls.

We have assumed numbers so that the nominal GDP in 2006 is the same in both cases. In case *a*, the nominal GDP is $120 for butter plus $430 for golf balls, for a total of $550. In case *b*, the nominal GDP is $105 for butter plus $445 for golf balls, for a total again of $550. However, the quantities of both goods are higher in case *b* than in case *a*. Thus, any sensible measure would indicate that the real GDP was higher in case *b* than in case *a*. Thus, the equality of the nominal GDPs is misleading. Identical figures on nominal GDP can conceal very different underlying differences in levels of production.

Economists solve the problem of changing price levels by constructing measures of real GDP. Until recently, the most common way to compute real GDP was to multiply each year's quantity of output of each good by the price of the good in a base year, such as 2000. Then all of these multiples were added to get the economy's aggregate real GDP. The resulting aggregate is called GDP in 2000 dollars (if 2000 is the base year). Or, sometimes, the result is called **GDP in constant dollars**, because we use prices (for the base year, 2000) that do not vary over time. As a contrast, the nominal GDP is sometimes called **GDP in current dollars**, because this calculation of GDP uses each good's price in the current year.

Since the prices from the base year (say, 2000) do not vary over time, the method just described gives a reasonable measure of changes over time in the overall level of production. That is, it gives a sensible measure of real GDP. However, a shortcoming of this approach is that it weights the outputs of the various goods by their prices in the base year (2000). These weights become less relevant over time as relative prices of goods change. The response of the Bureau of Economic Analysis (the BEA, a part of the U.S. Commerce Department) had been to make frequent shifts in the base year. However, a more accurate solution, called the chain-weighted method, was adopted in the mid 1990s to get a more reliable measure of real GDP. The resulting variable is called **chain-weighted real GDP**. This chain-weighted measure is the one publicized in the media, and is the one we shall use in this book to measure real GDP.

To illustrate the way chain-weighting works, we can again use our hypothetical data for a simple economy from Table 2.1. The method starts by computing the average price of each good for two adjacent years—2005 and 2006 in the table. For example, in scenario *a*, the average price of butter for 2005 and 2006 is $2.50 per pound. In scenario *b*, it is $1.75 per pound.

In each year—2005 and 2006 in the table—the quantities produced of each good are multiplied by the average prices for the two adjacent years. For example, in case *a*, the value of the butter produced in 2005 is $125 when calculated at the average price for 2005 and 2006, compared to $100 when the (lower) price for 2005 is used. For 2006 in case *a*, the value of the butter is $100 when computed at the average price, compared to $120 when the (higher) price for 2006 is used.

Using these average-price values, we sum the values of the goods produced in each to get the totals shown in Table 2.1. For example, for 2005 in case *a*, the total dollar value is $545, compared to $500 when we used prices for 2005. For 2006 in case *a*, the total dollar value is $510.6, compared to $550 when we used prices for 2006.

Next we compute the ratios of each of these totals to the totals for 2005. Thus, the ratios are 1.0 for the two cases (*a* and *b*) that apply to 2005. For 2006, the ratio is 0.937 in case *a* and 1.278 in case *b*.

To get chained real GDP on a 2005 base, we multiply the ratios just calculated by the nominal GDP ($500) for 2005. Thus, chained real GDP for 2005 on a 2005 base is the same as nominal GDP—$500 (for cases *a* and *b*). For 2006, chained real GDP on a 2005 base is $468.5 in case *a* and $639.0 in case *b*. Thus, although the nominal GDPs for 2006 are the same, the chained real GDP is substantially higher in case *b*. This result makes sense because the quantities of butter and golf balls are both higher in case *b* than in case *a*.

We can proceed in the same way for other years. For example, when we get data for 2007, we can calculate the ratio of the value of output in 2007 to that for 2006. These ratios are analogous to those shown for 2006 compared to 2005 in Table 2.1. We then want to express the results for 2007 on a 2005 base, so that all the chained values have the same base year. To do this, we multiply the ratio for 2007 compared to 2006 by the ratio for 2006 compared to 2005. This gives us the ratio of 2007 values to 2005 values. Finally, we multiply the last ratio by nominal GDP for 2005 to get the chain-weighted GDP for 2007 on a 2005 base. This procedure is called **chain-linking**. If we carry out this procedure from one year to the next, we end up with a time series for chain-weighted real GDP expressed in terms of a single base year.

In Table 2.1, the base year for chain-weighted real GDP is 2005. However, the actual base year used by the Bureau of Economic Analysis in the early 2000s was 2000. With the chain method, the choice of base year is not important. We use a single base year only to ensure that the real GDPs for each year are comparable.

We can use the results on real GDP to construct an index for the overall level of prices. In Table 2.1, where 2005 is the base year, we can think of the overall price level for 2005 as "100." This number is arbitrary; it just serves as a comparative position that can be related to price levels in other years.

For case *a* in 2006, the nominal GDP is \$550 and the chain-weighted real GDP on a 2005 base is \$468.5. We can think of an implicit price level that we are using to convert a dollar value—the nominal GDP of \$550—into a real value—the real GDP of \$468.5:

(nominal GDP)/(implicit price level) = real GDP.

That is, suppose that we start with a nominal GDP of \$11 trillion per year. Then, if we had just one type of good, we might have a price level of \$1.10 per good (where we measure goods in some physical unit). If we divide the nominal GDP by this price level, we get a measure of real GDP, which has units of goods per year: (\$11 trillion per year)/(\$1.10 per good) = 10 trillion goods per year.

If we rearrange the terms in the previous equation, we have

implicit price level = (nominal GDP)/(real GDP).

For example, for 2006 in case *a* in Table 2.1, we have

implicit price level = (550/468.5) = 1.17.

In contrast, for 2006 in case *b*, we have

implicit price level = (550/639.0) = 0.86.

The numbers 1.17 and 0.86 do not really mean anything as absolute magnitudes. However, they have meaning when compared with similarly calculated price levels for other years. As mentioned, the usual convention is to think of a price index that takes on the value 100 for the base year, which is 2005 in our example. When compared to this base, the price level for 2006 in case *a* is 1.17*100 = 117, whereas that in case *b* is 0.86*100 = 86. These values are shown in Table 2.1. The usual name for these price indexes is the **implicit GDP deflator** (on a 2005 base). That is, these values are the ones implicitly used to convert from nominal GDP to real GDP (on a 2005 base).

Although real GDP reveals a lot about an economy's overall performance, it is not a perfect measure of welfare. Some of the shortcomings of real GDP from a welfare standpoint are the following:

- The aggregate real GDP does not consider changes in the distribution of income.
- The calculated real GDP excludes most non-market goods. The exclusions include legal and illegal transactions in the "underground economy," as well as services that people perform in their homes. For example, if a person cares for his or her child at home, the real GDP excludes this service. But if the person hires someone to care for the child at home or at a day-care center, the real GDP would include the service.
- Real GDP assigns no value to leisure time.
- Measured real GDP does not consider changes in air and water quality, that is, pollution, except to the extent that this pollution affects the market value of output.

Despite these shortcomings, the real GDP tells us a lot about how an economy's standard of living changes over time. It also allows us to compare standards of living across countries. Thus, measured real GDP helps us to understand short-run economic fluctuations as well as long-term economic development.

II. Alternative Views of GDP—Expenditure, Income, and Production

We can think about the gross domestic product or GDP in three different ways. First, we can consider the expenditures on domestically produced goods and services by various groups—households, businesses, government, and foreigners. Second, we can calculate the incomes earned domestically in the production of goods and services—compensation of employees, rental income, corporate profits, and so on. Finally, we can measure the domestic production of goods and services by various industries—agriculture, manufacturing, wholesale and retail trade, and so on. An important point is that all three approaches will end up with the same totals for nominal and real GDP. To see this, we take up each approach in turn, beginning with the breakdown by type of expenditure.

A. Measuring GDP by expenditure

The national accounts divide GDP into four parts, depending on who or what buys the goods or services. The four sectors are households, businesses, all levels of government, and foreigners. Table 2.2 shows the details of this breakdown for 2003. The first column lists values in current (2003) dollars, and the second column uses chained real dollars in terms of the base year, 2000. The nominal GDP for 2003 of \$11.00 trillion corresponds to \$10.38 trillion in terms of 2000 dollars. If we make these calculations for other years, we can compare across years to see how real GDP changed over time.

The purchases of goods and services by households for consumption purposes are called **personal consumption expenditure**. This variable, like GDP, is a flow concept. Thus, nominal personal consumption expenditure has units of dollars per year. This spending accounts for the bulk of GDP. For example, Table 2.2 shows that, in 2003, the nominal personal consumption expenditure of $7.76 trillion was 71% of the nominal GDP of $11.00 trillion.

The national accounts distinguish purchases of consumer goods that are used up quickly, such as toothpaste and various services, from those that last for a substantial time, such as automobiles and furniture. The first group is called **consumer nondurables and services**, and the second is called **consumer durables**. An important point is that consumer durables yield a flow of services in the future, as well as when purchased. Table 2.2 shows the division of personal consumption expenditure among durable goods, nondurable goods, and services. Notice that, in 2003, the nominal spending on durables of $0.95 trillion accounted for only 12% of total nominal personal consumption expenditure.

The third column of Table 2.2 reports the components of GDP in chained 2000 dollars. For example, the nominal personal consumption expenditure of $7.76 trillion in 2003 corresponds to $7.36 trillion in 2000 dollars. If we apply this calculation to other years, we can compute the changes over time in real personal consumption expenditure or in the other real components of GDP. However, there are some difficulties in comparing real personal consumption expenditure with real GDP or the other real components of GDP. As already mentioned, the nominal personal consumption expenditure for 2003 was 71% of nominal GDP. However, a comparison of real personal consumption expenditure with real GDP depends on which base year one happens to use. That is because the comparison of real consumer expenditure with real GDP depends on the changes in relative prices that occurred between the base year (say 2000) and the comparison year, say 2003. In particular, the results depend on how prices of items contained in personal consumption expenditure changed compared to the prices of the other items that enter into GDP.

The second major category of GDP is **gross private domestic investment**. Investment, like personal consumption expenditure, is a flow variable, measured in dollars per year. The "fixed" part of gross private domestic investment comprises purchases by domestic businesses of new capital goods, such as factories and machinery. These capital goods are durables, which serve as inputs to production for many years. Thus, these goods are analogous to the consumer durables that we already mentioned. In fact, in the national accounts, an individual's purchase of a new home—which might be considered the ultimate consumer durable—is counted as part of fixed business investment, rather than personal consumption expenditure.

The total of gross private domestic investment is the sum of fixed investment and the net change in businesses' inventories of goods. In 2003, this net inventory change happened to be close to zero. The total of nominal gross private domestic investment was $1.67 trillion, which constituted 15% of nominal GDP.

One common error about national-income accounting arises because the spending on new physical capital is called "investment." This terminology differs from the concept of investment used in normal conversation, where investment refers to the allocation of financial assets among stocks, bonds, real estate, and so on. When economists refer to a business's investment, they mean the business's purchases of physical goods, such as a factory or machine.

Another point about investment concerns **depreciation**. The stock of capital goods is the outstanding quantity of goods in the form of factories, machinery, and so on.

Thus, the capital stock is a **stock variable**, measured as a quantity of goods. These capital goods wear out or depreciate over time. Hence, a part of gross investment merely replaces the old capital that has depreciated. Depreciation is a flow variable—the dollar value of the goods that wear out per year. Depreciation is comparable in units to GDP and gross private domestic investment.

The difference between gross private domestic investment and depreciation—called **net private domestic investment**—is the net change in the value of the stock of physical capital goods. The gross domestic product or GDP includes gross private domestic investment. If we replace this gross investment by net investment (by subtracting depreciation), we also subtract depreciation from GDP. The difference between GDP and depreciation is called **net domestic product** or NDP. The NDP is a useful concept because it measures GDP net of the spending needed to replace worn-out or depreciated capital goods.

The third component of GDP is government consumption and investment. This category includes consumption outlays (such as salaries of military personnel and public-school teachers), as well as public investment (such as purchases of new buildings). However, there is considerable arbitrariness about what the national accounts designate as government investment or government consumption. For example, purchases of military equipment, such as airplanes and tanks, are treated, despite their durability, as consumption. One important point is that the government sector includes all levels of government, whether federal, state, or local. Another point is that the category called government consumption and investment does not include transfers, such as payments to social-security retirees and welfare recipients. These transfers do not represent payments to individuals in exchange for currently produced goods or services. Therefore, these outlays do not appear in GDP. In 2003, nominal government consumption and investment was $2.08 trillion or 19% of nominal GDP.

Some of the goods and services produced domestically are exported to foreign users. These **exports** of goods and services must be added to domestic purchases to compute the economy's total domestic production (GDP). Foreigners also produce goods and services that are imported into the home country—for use by households, businesses, and government. These **imports** of goods and services must be subtracted from domestic purchases to calculate the economy's total production (GDP). The foreign component therefore appears in GDP as **net exports**: the difference between spending of foreigners on domestic production (exports) and spending by domestic residents on foreign production (imports). Net exports may be greater than zero or less than zero. Table 2.2 shows that, in 2003, net nominal exports were -$0.50 trillion or -4.5% of nominal GDP. The net export component breaks down into $1.05 trillion of exports (9.5% of GDP) less $1.54 trillion of imports (14% of GDP).

Economists often omit net exports to construct a macroeconomic model for a single economy. Then the model applies to a **closed economy**, which has no trade linkages to the rest of the world. In contrast, an economy that is linked through trade to the rest of the world is called an **open economy**. Some reasons for using a closed-economy model are the following:

- It simplifies the analysis.
- At least for the United States, exports and imports have been small compared to GDP, so that little error arises from ignoring international trade. This point was

reasonably persuasive in 1950 when exports and imports were each only 4% of GDP. However, the point is less convincing for 2003, when exports were 10% of GDP and imports were 14% of GDP.

- The world as a whole really is a closed economy, so we have to carry out a closed-economy analysis to assess the world economy.

We follow the closed-economy tradition of macroeconomics until chapter 17, which allows for international trade.

B. Measuring GDP by income

Another way to look at GDP is in terms of the income earned by various factors of production. This concept is called **national income**. To make clear the relation between production and income, we can think of a simple closed economy that has only two businesses. One, a mill, uses only labor to produce flour. The second, a bakery, uses flour and labor to produce bread. Bread is the only final product. Flour is the only intermediate product—it is used up entirely in the production of the final good, bread.

Income statements for the two businesses are in Table 2.3. The nominal GDP for this economy is the value of the final product, bread, of $600. This amount is also the revenue of the bakery. The income statement shows that the costs and profit for the bakery break down into $350 for flour, $200 for labor (for workers in the bakery), and $50 for profit (of the bakery). For the mill, the $350 of revenue goes for $250 of labor (for workers at the mill) and $100 for profit (of the mill). The national income equals the total labor income of $450 plus the total profit of $150, or $600. Thus, in this simple economy, the national income equals the GDP.

Notice that the GDP counts the value of the final product, bread, of $600, but does not count separately the value of the flour, $350. The flour is used up in the production of bread—that is, the $600 in bread sales already implicitly takes into account of the $350 cost of the intermediate good, flour. If we added the $350 in sales of flour to the $600 in sales of bread, we would double-count the contribution of the intermediate good, flour. To put it another way, the **value added** by the bakery is only $250—sales of $600 less payments for flour of $350. The value added by the mill is the full $350, because we assumed that the mill uses no intermediate goods. Therefore, if we combine the value added of $350 for the mill with the value added of $250 for the bakery, we get the GDP of $600. Hence, the GDP equals the sum of value added from all sectors. The national income in this simple economy equals the GDP and, therefore, also equals the sum of value added from all sectors.

Table 2.4 shows the breakdown of national income for the United States in 2003. The total nominal national income was $9.68 trillion. Although the method for computing national income is conceptually the same as that in Table 2.3, the U.S. economy includes additional forms of income. The largest part of U.S. national income was compensation of employees—$6.29 trillion or 65% of the total. This component is analogous to the labor income shown in Table 2.3.

Several parts of the U.S. national income in Table 2.4 represent income that accrues to capital. These amounts did not appear in Table 2.3 because we did not consider that the bakery and mill each have capital equipment, such as machinery, that contributes to the production of goods. In the U.S. national accounts, the categories of income from capital comprise rental income of persons, corporate profits, and net interest. The total of \$1.71 trillion represented 18% of national income.

The U.S. national income for 2003 also includes proprietors' income of \$0.83 trillion (9% of the total). This income represents payments to self-employed persons, including unincorporated businesses. This income represents a mix of payments to labor and capital.

Taxes on production—sales, excise, and value-added (or VAT)[2]—are included in market prices of goods. Therefore, these taxes on production appear in GDP, which is calculated from market values of output. The tax revenues are also part of government revenue—therefore, these revenues enter into national income as income of the government sector. Subsidies paid to producers by government amount to negative production taxes. Therefore, subsidies enter with a negative sign in national income. In 2003, the total of taxes on production less subsidies was \$0.75 trillion or 8% of national income.

National income also includes business's net transfers to households and government (\$0.08 trillion or 1% of national income). Finally, the national income includes the surplus of government enterprises. This surplus is analogous to corporate profits, except that the surplus accrues to the government sector. The amount in 2003 was only \$0.01 trillion or 0.1% of national income.

In the simplified economy of Table 2.3, GDP and national income were equal. In practice, divergences between GDP and national income reflect two main items: income receipts and payments involving the rest of the world and depreciation of capital stocks. We take up these two items in turn.

[2] The value-added tax is important in many countries but does not exist in the United States.

The U.S. GDP is the value of goods and services produced within the United States. The U.S. national income is the income received by all sectors residing in the United States. One source of divergence between GDP and national income is that residents of the United States receive income from the rest of the world. The main item is the income on capital (assets) owned by U.S. residents but located abroad. A secondary part is labor income of U.S. residents working abroad. The total of this "factor income from abroad" in 2003 was $0.33 trillion, as shown in Table 2.5. The counterpart to the U.S. factor income from abroad is the U.S. payments to factors located abroad. These payments are to capital (assets) located in the United States but owned by foreigners and to foreigners working in the United States. The total of these payments to the rest of the world in 2003 was 0.27 trillion. The **net factor income from abroad** is the difference between U.S. income receipts from the rest of the world and U.S. income payments to the rest of the world: $0.33 trillion less $0.27 trillion or $0.06 trillion. The addition of this amount to the GDP of $11.00 trillion yields the **gross national product (GNP)** of $11.06 trillion, as shown in Table 2.5. The GNP gives the total gross income to U.S. factors of production, whether located in the United States or abroad.

As mentioned, one part of U.S. GDP covers the depreciation of the fixed capital stock located in the United States. This depreciation does not show up as income for factors of production. In particular, depreciation is subtracted from gross business revenue to calculate corporate profits or proprietors' income. If we subtract from GNP the estimated depreciation of $1.35 trillion, we get the net national product (NNP) for 2003 of $9.71 trillion. Aside from a statistical discrepancy (-$0.03 trillion in 2003), the NNP corresponds to national income. Thus, Table 2.5 shows how we get to the national income of \$9.68 trillion, the number that we saw before in Table 2.4.

We can also calculate the income that households receive directly, a concept called **personal income**. The route from national income to personal income involves a number of adjustments. First, only a portion of corporate profits are paid out as dividends to households. Second, personal income excludes contributions for government social-insurance programs, because individuals do not receive these contributions directly as income. Other adjustments involve transfer payments and the surplus of government enterprises. Table 2.5 lists the various items. The personal income for 2003 of \$9.16 trillion turned out to be \$0.52 trillion less than national income.

We can also compute the income that households have left after paying personal taxes, which include individual income taxes and property taxes. (Some other taxes—those on production and contributions for social insurance—were already deducted to calculate personal income.) Personal income after taxes is called **disposable personal income**. In 2003, personal taxes were \$1.00 trillion. Deducting this amount from the personal income of \$9.16 trillion leads to the disposable personal income of \$8.16 trillion, as shown in Table 2.5.

C. Measuring GDP by production

We can also break down national income in accordance with the sectors of production that generate the income. Table 2.6 shows this breakdown for the United States in 2003.

The total national income of $9.40 trillion[3] breaks down into $9.34 trillion from domestic industries and $0.06 trillion from the rest of the world, that is, the net factor income from abroad. For the domestic industries, $8.16 trillion or 87% comes from the private sector and $1.18 trillion or 13% from government (federal, state, and local).

The table shows how the $8.16 trillion from private industries divides into 14 sectors. The largest shares are 21% in finance, insurance, and real estate; 15% in professional and business services; 14% in manufacturing; 10% in education, health care, and social assistance; 9% in retail trade; 7% in wholesale trade; 6% in construction; 4% in arts, entertainment, recreation, accommodation, and food services; 4% in information; and 3% in transportation. Note that agriculture and mining together constitute only 2% of the total.

D. Seasonal adjustment

Data on GDP and its components are available for the United States and most other countries on a quarterly basis. These data allow us to study economic fluctuations at a quarterly frequency. However, one problem with the raw data is that they include sizeable systematic variation due to seasonal factors. The typical pattern is that real GDP rises during a calendar year and reaches a peak in the fourth quarter (October-December). Then real GDP usually falls sharply in the first quarter of the next year (January-March) before rebounding from the second to the fourth quarters.

The seasonal fluctuations in real GDP and other macroeconomic variables reflect the influences of weather and holidays (notably the Christmas period and summer

[3] This total is less by $0.28 trillion than the national income of $9.68 trillion shown in Tables 2.4 and 2.5 because of a differing treatment of depreciation.

vacations). For most purposes, we want to use the national-accounts data to study economic fluctuations that reflect factors other than normal seasonal patterns. For this reason, the Bureau of Economic Analysis (BEA) adjusts real GDP and its components to filter out the typical seasonal variation. Variables adjusted this way are called **seasonally-adjusted data**. The national-accounts information reported in the news media and used for most macroeconomic analyses come in this seasonally-adjusted form. We shall also use seasonally-adjusted quarterly data in this book to analyze economic fluctuations.

Seasonal adjustments apply also to many of the monthly variables reported in the news media and used for macroeconomic analyses. These variables include employment and unemployment, labor earnings, industrial production, retail sales, and the consumer price index.[4] When we discuss these monthly variables in this book, we refer to seasonally-adjusted data.

Gross state product for U.S. states

We have focused on the overall U.S. GDP. It is also possible to break down GDP into amounts produced within each of the 51 U.S. states (including the District of Columbia). The value of the gross output produced in a state is called **gross state product** (GSP). Table 2.7 shows how the total nominal U.S. GDP of $10.1 trillion in 2001 broke down by state. Note that California

[4] The seasonal variation in the CPI turns out to be minor but that in the other variables is substantial. Seasonal variation is not detectable in various interest rates, and these variables are not seasonally adjusted.

contributed 13.4% of U.S. GDP, New York 8.2%, and Texas 7.5%. At the low end, Montana, South Dakota, and Vermont each had only 0.2% of U.S. GDP.

Table 2.7
Gross State Product by U.S. State in 2001

State	Gross state product (billions of dollars)	% of total U.S. GDP	State	Gross state product (billions of dollars)	% of total U.S. GDP
U.S.	10137	100.0	Mississippi	67	0.7
Alabama	121	1.2	Missouri	181	1.8
Alaska	29	0.3	Montana	23	0.2
Arizona	161	1.6	Nebraska	57	0.6
Arkansas	68	0.7	Nevada	79	0.8
California	1359	13.4	New Hampshire	47	0.5
Colorado	174	1.7	New Jersey	365	3.6
Connecticut	166	1.6	New Mexico	55	0.5
Delaware	41	0.4	New York	826	8.2
D.C.	64	0.6	North Carolina	276	2.7
Florida	491	4.8	North Dakota	19	0.2
Georgia	300	3.0	Ohio	374	3.7
Hawaii	44	0.4	Oklahoma	94	0.9
Idaho	37	0.4	Oregon	120	1.2
Illinois	476	4.7	Pennsylvania	408	4.0
Indiana	190	1.9	Rhode Island	37	0.4
Iowa	91	0.9	South Carolina	115	1.1
Kansas	87	0.9	South Dakota	24	0.2
Kentucky	120	1.2	Tennessee	183	1.8
Louisiana	149	1.5	Texas	764	7.5
Maine	37	0.4	Utah	70	0.7
Maryland	195	1.9	Vermont	19	0.2
Massachusetts	288	2.8	Virginia	273	2.7
Michigan	320	3.2	Washington	223	2.2
Minnesota	188	1.9	West Virginia	42	0.4

Source: Bureau of Economic Analysis.

III. Prices

We already discussed how the computation of chained real GDP generates an implicit price deflator for the GDP. The resulting series gives us a good measure of the overall price level. That is, we get a price index that matches up with the economy's overall market basket of goods produced domestically. We can also use this approach to get implicit price deflators for the various components of GDP. For example, we have a deflator for personal consumption expenditure, one for gross private domestic investment, and so on.[5]

In addition to these implicit price deflators, we have broad price indexes that the Bureau of Labor Statistics (BLS) computes directly. The most important examples are the **consumer price index** (**CPI**) and the **producer price index** (**PPI**), which is also called the **wholesale price index**.

The main CPI series comes from a monthly survey of prices of goods and services in 87 urban areas. Data are collected on roughly 80,000 items from 23,000 retail and service establishments. The CPI also includes data on rents, which come from a survey of 50,000 landlords or tenants. The index that receives the most attention applies to urban consumers, estimated to cover 87% of the overall U.S. population. The CPI is a weighted average of individual prices, where the current weights reflect expenditure shares found in the Consumer Expenditure Survey of over 30,000 individuals and families in 1993-95. These weights remain fixed from month to month, until a new survey is taken.[6]

---

[5] However, the deflator for government consumption and investment is not very useful. Since most of government output is not sold on markets, this price deflator reflects arbitrary assumptions about costs of providing public services. In particular, the main assumption is that the productivity of government employees does not vary over time.

[6] The CPI, which weights individual prices by the importance of goods in a prior year (such as 1993-95), is an example of a Laspeyres index of prices. In contrast, a price index that weights by the importance of goods in the current year is called a Paasche index. The old-style implicit GDP deflator, which weighted

The CPI in the early 2000s had a base of 1982-84 = 100. Thus, the CPI for August 2004 of 189.4 meant that the price of the average item rose by 89.4% from mid 1983 to August 2004. This cumulative increase in the price level corresponded to an average growth rate of the price level by 3.0% per year over the 21 years since mid 1983. That is, the **inflation rate** measured by the CPI was 3.0% per year. In contrast, the inflation rate computed from the implicit GDP deflator over the same period was 2.4% per year. As discussed in the nearby box, the higher CPI inflation rate—by 0.6% per year—probably reflects an upward bias created by the fixed weights in the CPI market basket.

The PPI is computed in a manner conceptually similar to that for the CPI. However, the PPI does not cover services and primarily includes goods that are raw materials and semi-finished products. Each month the PPI survey collects about 100,000 prices from about 30,000 businesses. One shortcoming of the PPI is that it is too narrow a concept to reflect the general level of prices in an economy.

Problems with the Consumer Price Index

The CPI receives a lot of attention because it provides monthly information on the prices of a broad market basket of goods and services. Part of the attention arises because some public and private contracts index nominal payments to the CPI. For example, benefits paid under social security, payments

current expenditure in accordance with prices from a prior base year, turns out to be a Paasche index of prices. However, the modern version of the implicit GDP deflator is a chain-linked index, where the weights effectively change with each observation. The PPI is another example of a Laspeyres index of prices.

made on the U.S. Treasury's inflation-protected securities, and features such as bracket limits in the U.S. individual income tax adjust automatically for variations in the CPI.

Many economists think that the reported changes in the CPI overstate inflation and, hence, that the automatic adjustments of social-security benefits and some other payments have been too large to keep the outlays fixed in real terms. Naturally, this assessment is controversial, because any repairs would have significant consequences for transfer payments, tax collections, and so on. The idea that changes in the CPI seriously overstate inflation was expressed in 1996 by the President's Commission on the Consumer Price Index (see Michael Boskin, et al [1996]). The conclusion they reached was that the growth rate of the CPI exaggerated inflation on average by over one percentage point per year.

One reason for the overstatement of inflation is called substitution bias. The idea is that changes in supply conditions shift the relative prices of various goods and services, and households respond by shifting expenditure toward the goods and services that have become relatively cheaper. However, because the weights in the CPI are fixed for long intervals, the index responds only with a long lag to changes in the pattern of purchases. In particular, the CPI fails to give increasing weight to the cheaper items that tend to become more important in the typical household's market basket. This problem is conceptually easy to fix by shifting to the chain-weighting approach that we described before for the calculation of the implicit GDP deflator. This deflator is free of substitution bias because the weights change nearly continuously over time.

The BLS has, in fact, been experimenting with a chain-weighted measure of the CPI. The new series, available since January 2000, showed an annual inflation rate of 2.07% per year from January 2000 to August 2004, compared to 2.52% per year for the standard CPI over the same period. The inflation rate from the implicit GDP deflator (a chain-weighted index) over the same period was 2.00% per year—very close to the chain-weighted CPI. These comparisons suggest that the substitution bias in the standard CPI led to an overstatement of inflation by 0.4-0.5% per year.

Another, more challenging, problem with all of the price indexes—the implicit GDP deflator as well as the CPI—involves quality change. Despite attempts to measure improvements in quality, these changes tend to be underestimated. Therefore, some of the price increases that are recorded as inflation should actually be viewed as increases in money spent to get better quality products. A full accounting for quality improvements would therefore lower the inflation rate. Some improved measurement has been made for goods such as automobiles, computers, houses, and television sets. Interesting proposals for measuring quality change have also been offered in the medical area, where technical advances that save lives or improve the quality of life tend to be labeled as inflation.

Another problem is that the various price indexes do not consider the effective reductions in the price level that occur when new products are introduced. For example, when personal computers or DVD players were introduced, households were made better off for a given dollar income—even if the new goods were initially “expensive.” The same idea applies to the invention of new prescription drugs, even if the prices of these drugs are “high” at the outset. The creation of useful new products tends to raise households’ real income or, equivalently, lower the effective price level. Thus, a proper accounting for new products would lower the average inflation rate. Equivalently, the economy’s real economic growth would look stronger if the effects of new products were properly considered.

Table 2.1

The Calculation of Nominal and Real GDP—A Simple Example

	2005a	2005b	2006a	2006b
Prices				
butter	$2.00 per pound	$2.00 per pound	$3.00 per pound	$1.50 per pound
golf balls	$1.00 per ball	$1.00 per ball	$1.10 per ball	$0.89 per ball
Quantities				
butter	50 pounds	50 pounds	40 pounds	70 pounds
golf balls	400 balls	400 balls	391 balls	500 balls
Nominal market values				
butter	100	100	120	105
golf balls	400	400	430	445
nominal GDP	500	500	550	550
2005-06 average price				
butter	$2.50 per pound	$1.75 per pound	$2.50 per pound	$1.75 per pound
golf balls	$1.05 per ball	$0.945 per ball	$1.05 per ball	$0.945 per ball
Market values at 2005-06 average prices				
butter	125.0	87.5	100.0	122.5
golf balls	420.0	378.0	410.6	472.5
total	545.0	465.5	510.6	595.0
ratio to 2005 total	1.0	1.0	0.937	1.278
Chained real GDP, 2005 base	500.0	500.0	468.5	639.0
Implicit GDP deflator, 2005 base	100	100	117	86

Table 2.2

Expenditure Components of the U.S. Gross Domestic Product in 2003

Category of Expenditure	Trillions of dollars	Trillions of chained 2000 dollars
Gross domestic product	**11.00**	**10.38**
Personal consumption expenditure	**7.76**	**7.36**
durable goods	0.95	1.03
nondurable goods	2.20	2.11
services	4.61	4.22
Gross private domestic investment	**1.67**	**1.63**
fixed investment	1.67	1.63
nonresidential	1.09	1.11
residential	0.57	0.51
change in business inventories	0.00	0.00
Government consumption & investment	**2.08**	**1.91**
federal	0.75	0.69
state and local	1.32	1.22
Net exports of goods and services	**-0.50**	**-0.52**
exports	1.05	1.03
imports	1.54	1.55

Source: Bureau of Economic Analysis.

Table 2.3

Hypothetical Data for Calculation of National Income

Type of revenue	Amount	Type of cost or profit	Amount
Bakery (produces final good)			
Sale of bread	$600	Labor	$200
		Flour	350
		Profit	50
		Total cost & profit	$600
Mill (produces intermediate good)			
Sale of flour	$350	Labor	$250
		Profit	100
		Total cost & profit	$350

Table 2.4

U.S. National Income by Type in 2003

Type of income	Trillions of dollars
National income	**9.68**
Compensation of employees	6.29
Proprietors' income	0.83
Rental income of persons	0.15
Corporate profits	1.02
Net interest	0.54
Taxes on production	0.80
less: subsidies	(0.05)
Business transfers	0.08
Surplus of government enterprises	0.01

Source: Bureau of Economic Analysis.

Table 2.5

Relations between U.S. GDP and Income in 2003

Category of product or income	**Trillions of dollars**
Gross domestic product (GDP)	**11.00**
plus: income receipts from rest of world	0.33
less: income payments to rest of world	(0.27)
equals: Gross national product (GNP)	**11.06**
less: depreciation of capital stock	(1.35)
equals: Net national product (NNP)	**9.71**
less: statistical discrepancy	(0.03)
equals: National income	**9.68**
less: corporate profits, taxes on production, contributions for social insurance, net interest, business transfers, surplus of government enterprises	(3.18)
plus: personal income receipts on assets, personal transfer payments	2.66
equals: Personal income	**9.16**
less: personal taxes	1.00
equals: Disposable personal income	**8.16**

Source: Bureau of Economic Analysis.

Table 2.6

U.S. National Income by Sector in 2003

Sector of production	Trillions of dollars
National income*	**9.40**
Domestic industries	**9.34**
Private industries	**8.16**
agriculture, forestry, fishing, hunting	0.08
mining	0.09
utilities	0.16
construction	0.48
manufacturing	1.11
wholesale trade	0.57
retail trade	0.73
transportation, warehousing	0.26
information	0.31
finance, insurance, real estate, rental, leasing	1.74
professional and business services	1.23
education, health care, social assistance	0.83
arts, entertainment, recreation, accommodation, food services	0.34
other private services	0.24
Government	**1.18**
Net factor income from abroad	**0.06**

Source: Bureau of Economic Analysis.

*This total is less by $0.28 trillion than the national income of $9.68 trillion shown in Tables 2.4 and 2.5 because of a differing treatment of depreciation.

January 12, 2005

Chapter 3

Facts on Growth; The Solow Growth Model

In 2000, the real gross domestic product or real GDP per person in the United States was $32,600, valued using 1995 U.S. dollars. This high level of output and income per person meant that the typical U.S. resident had a high **standard of living**. Most adults had their own home, at least one car, several television sets, education at least through high school, and a level of health that meant a life expectancy at birth of nearly 80 years. Similar standards of living were enjoyed in most Western European countries—including Germany, France, the United Kingdom, and Italy—and a few other places, such as Canada, Australia, New Zealand, Japan, Singapore, and Hong Kong.

In contrast, the residents of most other countries were not nearly as well off in 2000. For example, the real GDP per person was $8900 in Mexico, $4000 in China, $2500 in India, and $1000 in Nigeria, the largest country in Africa.[1] These lower real GDPs and real incomes per person meant lower standards of living. The typical Mexican did reasonably well in food, shelter, and basic healthcare but could not afford the range and quality of consumer goods available to most Americans. Even more seriously, the typical Nigerian had concerns about nutrition and housing and faced a life expectancy at birth of less than 50 years.

The question is how can countries with low real GDP per person catch up to the high levels enjoyed by the United States and other rich countries? The only answer is

[1] These GDP numbers adjust for purchasing-power differences across countries. The data are from Robert Summers and Alan Heston (2002).

that the poorer countries must have high rates of economic growth over long periods—20 years, 40 years, and longer. For example, Table 3.1 shows the level of real GDP per person that China would attain in 2020 depending on its growth rate of real GDP per person from 2000 to 2020. It would take a growth rate of 10% per year—an unprecedented accomplishment for 20 years—for China's real GDP per person in 2020 to reach $30000, nearly the U.S. level in 2000.[2] In contrast, if China's real GDP per person grew at 2% per year, its level of real GDP per person in 2020 would be only $6000, 18% of the U.S. level in 2000. Thus, differences in rates of economic growth, when sustained for 20 years or longer, make an enormous difference in standards of living, as represented by levels of real GDP per person.

Table 3.1
Economic Growth and China's real GDP per person in 2020[3]

Growth rate of real GDP per person from 2000 to 2020	Real GDP per person in 2020
2% per year	6000
5% per year	11000
10% per year	30000

Once we appreciate the significance of sustained economic growth, the question is what can we—or our governments—do to make the rate of economic growth higher? The importance of this question inspired Bob Lucas (1988) to say: "Is there some action a government of India could take that would lead the Indian economy to grow like

[2] Since U.S. real GDP per person will likely also be growing, China's real GDP per person in 2020 would still be substantially lower than the U.S. real GDP per person in 2020.

[3] We can calculate the level of real GDP per person in 2020 as follows. Start with the natural logarithm of real GDP per person in 2000: ln(4000) = 8.294. Then multiply the number of years, 20, by the growth rate, for example 0.02 if the growth rate is 2% per year: 20 x 0.02 = 0.40. Add this to 8.294 to get 8.694. Then take the exponential of 8.694 to get the answer, 5970.

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."[4] Questions like these—especially the challenge to develop policies that promote economic growth—motivate the study that we begin in this chapter and continue in the following two chapters.

We start by outlining key facts about economic growth, first for a large number of countries since 1960 and, second, for the United States and other rich countries for over a century. These observations bring out patterns that we need to understand in order to design policies that promote economic growth. As a way to gain this understanding, we construct a model of economic growth—the Solow model. In the next two chapters, we extend this model and see how these extensions relate to patterns of economic growth and to Lucas's policy challenge.

I. Economic Growth around the World, 1960-2000

We mentioned that, in 2000, the real GDP per person in the United States was \$32,600, but that the real GDP per person in most other countries was substantially lower. We can illustrate the differences by graphing the real GDP per person in 2000 for a large number of countries. When we do, we get the distribution shown in Figure 3.1. The horizontal axis plots real GDP per person, and the vertical axis shows the number of

[4] When Lucas wrote these words in the mid 1980s, India had been growing more slowly than Egypt and Indonesia. The growth rates of real GDP per person from 1960 to 1980 were 3.2% per year in Egypt, 3.9% in Indonesia, and 1.5% in India. However, India did manage to surpass the other two countries in terms of growth rates from 1980 to 2000: the growth rates of real GDP per person were 1.8% per year for Egypt, 3.5% for Indonesia, and 3.6% for India. Thus, the Indian government may have met Lucas's challenge. However, the growth rates also fell in Egypt and Indonesia.

countries that had each real GDP per person. The graph summarizes the data for 147 countries, and representative countries are labeled for each bar.

The United States was only the second richest country in 2000—the number one spot went to Luxembourg, a very small country, with a real GDP per person of over $44,000. More generally, the top positions were dominated by the long-term members of the rich countries' club, which is known as the Organization for Economic Cooperation and Development or OECD.[5] This elite group includes most of Western Europe, the United States, Canada, Australia, New Zealand, and Japan. Overall, 21 of the richest 25 countries in 2000 were OECD members. The other four were Singapore (rank 3), Hong Kong (rank 4, if viewed as an economy separate from China), Cyprus (22), and Taiwan (23).[6]

The poorest country in Figure 3.1 is Tanzania, a sub-Saharan African country, with a real GDP per person of $570, again in 1995 U.S. dollars.[7] Therefore, the richest country (Luxembourg) had a real GDP per person that was 78 times greater than the poorest country. If we ignore Luxembourg because of its small size and compare instead with the United States, we find that the United States had a real GDP per person that was 57 times as large as Tanzania's.

Economists use the term **poverty** to identify low standards of living. A person or family living in poverty has difficulty in sustaining the basic necessities of life—food,

---

[5] The OECD was formed in December 1960 with 20 member states: Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States. Four other countries joined by the end of the 1960s—Australia, Finland, Japan, and New Zealand. By 2002, the OECD had added six more countries to attain a membership of 30—the six recent entrants were Czech Republic, Hungary, Mexico, Poland, Slovak Republic, and South Korea.

[6] One curious aspect of Taiwan is that it is not recognized as a state by the International Monetary Fund or the World Bank; hence, statistics on Taiwan have traditionally been omitted from these standard sources.

[7] The democratic republic of Congo—the former Zaire—would probably occupy the lowest position except for missing data.

clothing, shelter, and health—and can only dream about automobiles and television sets. We can identify poverty with low incomes of individuals and families. According to one commonly accepted definition from the United Nations and the World Bank, a person was living in poverty in 2000 if his or her annual income was less than $1000 and was in extreme poverty if this income was less than $500.[8] The number of persons in poverty in a country depends on two things. One is the way that the country's income is distributed among persons, for example, the total income might be distributed nearly evenly or a small fraction of the population might receive most of the income. The second is the country's average real income, which we can measure by real GDP per person. If this average is very low, the typical resident will be living in poverty even if income is distributed evenly.

In practice, the second factor—a country's real GDP per person—is the most important determinant of the number of people living in poverty. Countries with very low real GDP per person are the ones in which a large fraction of the population lives in poverty. Therefore, the data that underlie Figure 3.1 tell us that world poverty in 2000 was dominated by sub-Saharan Africa—an amazing 23 of the lowest 25 values of real GDP per person were in this region. The two other countries in this poorest group were Yemen (14th from the bottom) and Cambodia (24th).

The values of real GDP per person in 2000 give us a snapshot of standards of living at a point in time. The reason that the rich countries, such as the OECD members, got to be rich in 2000 is that their level of real GDP per person rose for a long period. Similarly, the poor countries in 2000—especially in sub-Saharan Africa—did not grow.

[8] The original definitions for the mid 1980s were $2 per day for poverty and $1 per day for extreme poverty. The annual values for 2000 adjust the original numbers for changes in price levels.

In fact, as we shall see, many of these growth rates were negative, so that real GDP per person fell over time.

To measure economic growth, we have to compare the levels of real GDP per person in 2000 with those from earlier years. Figure 3.2 begins this comparison by showing a graph for real GDP per person forty years earlier, in 1960.[9] This graph is similar to the one in Figure 3.1. The horizontal axis again shows real GDP per person, still using 1995 U.S. dollars. The vertical axis shows the number of countries that had each real GDP per person in 1960. The total number of countries is only 113, given the availability of data for 1960.

In 1960, Switzerland was at the top with a real GDP per person of $15,200, and the United States was still second, at $12,800. The top 25 was again dominated by the long-term members of the OECD—in 1960, 19 of the richest 25 countries were OECD members.[10] One difference from 2000 was that no Asian countries were in the top 25 in 1960. Several Latin American countries (Argentina, Uruguay, Venezuela, and Barbados) were in the top group, but none of these remained in 2000. Israel and South Africa were also in this group in 1960 but not in 2000 (where Israel was 27th and South Africa 54th).

The low end for real GDP per person was dominated somewhat less by sub-Saharan Africa in 1960 than in 2000. The poorest country in 1960 was still Tanzania at $450, but "only" 17 of the 25 countries with the lowest real GDP per person were in sub-Saharan Africa. Many of the poorest 25 in 1960 were in Asia—China, Indonesia, Nepal, Pakistan, India, and Bangladesh. All of these countries except Nepal grew rapidly enough over the next 40 years to escape the bottom category. In fact, the high growth in

[9] For international data on gross domestic product, 1960 is the earliest year in which reasonably accurate information is available for most countries.
[10] Japan and Spain were just below the top 25 in 1960 but made the top group in 2000.

Asia and the low growth in sub-Saharan Africa from 1960 to 2000 were major parts of the story about world standards of living in 2000. In a nearby boxed section, we discuss how these developments affected world poverty.

In 1960, the richest country (Switzerland) had a real GDP per person that was 34 times that of the poorest country (Tanzania). This spread was lower than the one we found for 2000, where the U.S. real GDP per person was 57 times the one in Tanzania.

If we compare the levels of real GDP per person in 2000 and 1960 for each country, we can compute the country's growth rate of real GDP per person over the 40 years.[11] Figure 3.3 shows the distribution of these growth rates for the 111 countries with the necessary data. The construction of this graph is similar to that for Figures 3.1 and 3.2. However, the horizontal axis now shows the growth rate of real GDP per person from 1960 to 2000, and the vertical axis shows the number of countries that had each growth rate.

The average growth rate of real GDP per person from 1960 to 2000 for the 111 countries was 1.8% per year. The fastest growing country was Taiwan, with a rate of 6.4%. More generally, many of the fast growers from 1960 to 2000—9 out of 13—came from East Asia. Aside from Taiwan, the East Asian countries with fast growth were Singapore, South Korea, Hong Kong, Thailand, China, Japan (which grew rapidly only up to the early 1970s), Malaysia, and Indonesia. Some long-term OECD members were among the top-20 for economic growth: Ireland, Portugal, Spain, Greece, Luxembourg, and Norway. The other fast-growing countries were Botswana (the star performer of sub-Saharan Africa), Cyprus, Romania, Mauritius, and Cape Verde.

[11] The easiest way to compute the growth rate of real GDP per person from 1960 to 2000 is to calculate (1/40)*ln(real GDP per person in 2000/real GDP per person in 1960), where ln is the natural logarithm.

At the bottom end, 18 of the 20 worst performers for economic growth from 1960 to 2000 were in sub-Saharan Africa.[12] Among these 18 African countries, 14 experienced negative growth of real GDP per person, with Zambia the worst at -1.8%. Thus, the reason for low levels of real GDP per person in 2000 is partly that countries in this region started off badly in 1960 (around the times of independence for most of the countries) but, even more so, that they performed so poorly in terms of the growth of real GDP per person from 1960 to 2000.

One thing we have learned is that the poorest countries in 2000—especially in sub-Saharan Africa—were poor because they grew at low or negative rates since 1960. Thus, to go further, we have to understand why these countries failed to grow at higher rates.

We have also learned that a group of countries in East Asia grew at high rates from 1960 to 2000. This strong growth enabled these countries to move up from low levels of real GDP per person in 1960 to much higher levels in 2000. To understand this change, we have to understand why these countries grew at high rates.

To appreciate the high levels of real GDP per person in the OECD countries in 2000, we have to look at data before 1960. That is, these countries were rich in 2000 partly because they grew from 1960 to 2000 but, even more so, because they were already rich in 1960. To get an appreciation of these longer term developments, we turn to information for the United States.

[12] The other two of the lowest 20 growers for 1960-2000 were Nicaragua (-0.6%) and Venezuela (0.2%).

World poverty and income inequality

We mentioned that poverty refers to a minimally acceptable standard of living. For example, the standard might be based on the amount of real income needed to sustain life through the provision of food, shelter, clothing, and healthcare. Thus, we might designate a standard as a specified amount of real income. We mentioned that two common standards—used by international organizations such as the United Nations and the World Bank—are $1 or $2 per day, expressed in 1985 dollars. In 2000, these amounts translated into annual incomes per person of about $500 and $1000.

The term **inequality** is often used interchangeably with poverty but is actually very different. Inequality involves the distribution of income across individuals at a point in time within a country or around the world. Economists have used several measures of inequality—examples are the fractions of a country's total income that go to persons in the lowest and highest fifths of the distribution. If income were equally distributed, both of these numbers would be 20%. In practice, the distribution of income is far from equal—for 73 countries with data around 1990, the average of the income going to the lowest fifth was 6.6%, whereas that for the highest fifth was 45%.[13] As examples, the numbers were 6.5% and 39% for the United States (about average inequality based on the

[13] These numbers refer to estimates of income distributions for incomes measured net of taxes.

lowest fifth and lower than average inequality based on the highest fifth), 9.4% and 28% for Canada (a country with relatively little inequality), 4.4% and 60% for Brazil (a country with a lot of inequality), and 7.8% and 41% for the United Kingdom.

For a given total income—for example, for a given real GDP per person—the degree of inequality determines the fraction of the population that falls below a poverty threshold, such as $1 or $2 per day. Unless the real GDP per person is very low, more inequality means that a higher fraction of the population falls below each of these poverty lines.[14] However, when the total income changes—for example, when real GDP per person rises—inequality and poverty behave very differently.

To see this, suppose that everyone's real income were to double. In this case, inequality would not change—for example, if the lowest fifth of the distribution started with 6% of the total income, the lowest fifth would still have 6% after everyone's income had doubled. In contrast, poverty would fall sharply if everyone's income doubled—because more people's incomes would exceed $1 or $2 per day. If we think that a person's welfare depends on his or her real

[14] Suppose, instead, that a country's real GDP per person is very low, so that income per person is below the poverty level of $1 per day. If everyone has the same income, 100% of the population would be living in poverty. In contrast, if income is unequally distributed, some fraction of the population would be above the poverty level. Therefore, in this case, an increase in inequality leads to a smaller fraction of the population living in poverty.

income, rather than the income measured relative to that of other persons, then poverty is more meaningful than inequality as a measure of welfare.

Xavier Sala-i-Martin showed in a recent study ("The World Distribution of Income, Estimated from Individual Country Distributions," Columbia University, April 2002) that world economic growth led to a dramatic fall in poverty from 1970 to 1998. The estimated number of people below the $1-a-day standard fell from 550 million, or 17% of the world's population, in 1970 to 350 million, or 7% of the population, in 1998. For the $2-a-day standard, the decline was from 1.3 billion, or 41% of world population, to 970 million, or 19% of the population.[15]

Figures 3.4a and 3.4b show how these changes occurred. The upper graph has the distribution of income for the world's people in 1970. The horizontal axis plots the log of real income, and the vertical axis shows the number of people who had each level of income. The vertical lines marked $1 and $2 show the income levels that correspond to the two poverty standards. Consider the area colored @ in the graph—it lies below the upper curve and to the left of the $1 poverty line. To find the fraction of the world's population that had incomes below $1 per day, we take the ratio of this area to the total area under the upper

[15] The fraction of the population living in poverty is called the **poverty rate**, whereas the number of people living in poverty is called the **poverty headcount**. The decrease in the poverty rate from 1970 to 1998 was so sharp that poverty headcounts also decreased, despite the substantial rise in world population (from 3.2 billion to 5.1 billion for the roughly 90% of the world's population included in Sala-i-Martin's study).

curve. The result is 17%. Similarly, to find the fraction below $2 per day, we take the ratio of the area colored @ to the total area. In this case, we get 41%.

World economic growth from 1970 to 1998 led to a shift from the upper curve in Figure 3.4a to the one in Figure 3.4b. Notice that the whole distribution shifted to the right—because larger proportions of the world's people had higher real incomes. Hence, the fractions of the world's population that had incomes below the $1 and $2 per day poverty lines were much smaller in 1998 than in 1970. The percentages for 1998 turned out to be 7% and 19%, respectively. This sharp decline in poverty rates shows the dramatic progress that can occur over three decades because of economic growth.

The graphs also show how some of the world's largest countries fared from 1970 to 1998. For poverty, the biggest stories are China and India, which accounted for 38% of world population in 1998. In 1970, many residents of China and India and other Asian countries were below the poverty lines—Asia accounted overall for 76% of persons below $1 per day and 85% of those below $2 per day. However, Figures 4a and 4b show that the income distribution curves for China and India (and also Indonesia, another large Asian country) shifted dramatically to the right from 1970 to 1998. This change reflected the strong Asian economic growth, particularly since the late 1970s for China and early 1980s for India. Consequently, by 1998, Asia accounted for only 15% of persons below the $1 per day poverty line and 49% of those below the $2 per day line.

We also know that recent decades saw very low economic growth in Africa. Consequently, the poverty numbers soared. In 1970, Africa accounted for only 11% of persons below $1 per day and also 11% of those below $2 per day. However, in 1998, Africa accounted for 67% of persons below $1 per day and 38% of those below $2 per day. Thus, particularly if we look at the $1 per day measure, poverty shifted from being primarily a problem in Asia to one principally in Africa.

For world inequality, the results are more complicated. We can think of the changes in two parts. The first is within countries and the second is across countries. Inequality rose from 1970 to 1998 within several large countries, including the United States, United Kingdom, Australia, and China. However, Sala-i-Martin showed that this change within countries had only a minor effect on the inequality across persons in the entire world.

The second factor is the dispersion of average incomes across countries. Our first guess is that this spread increased from 1970 to 1998. For example, we know from Figures 3.1 and 3.2 that the ratio of the highest real GDPs per person (concentrated in the OECD countries) to the lowest (primarily in sub-Saharan Africa) rose from 1960 to 2000. However, since world inequality involves numbers of people, rather than numbers of countries, we have to give more weight to the larger countries. Thus, the income changes in the larger countries, notably China and India, matter a lot more than the changes in the smaller countries. Since China and India had very low average incomes in 1970, their strong economic growth from 1970 to 1998 contributed a lot to a reduction of world income inequality. It turns out that this force dominates the others and leads to a decrease in standard measures of world inequality from 1970 to 1998.

II. Long-Term Growth in the United States and other Rich Countries

For the United States and other OECD countries, the main reason for today's high real GDP per person is that these countries already had high real GDP per person in 1960. Therefore, to understand the source of today's prosperity, we have to take a long-term view that starts well before 1960.

If we go back more than a century, we find that U.S. real GDP per person in 1869 was $2170, measured in 1995 U.S. dollars, compared to $33,500 in 2003.[16] Therefore, real GDP per person in 2003 was 15.5 times that in 1869. An increase in real GDP per person and, hence, in the typical person's real income by a factor of 15 makes an enormous difference in the standard of living. In 2003, unlike in 1869, the typical U.S. family not only owned a comfortable home and had ample food and clothing but also possessed many things not even imagined in 1869, such as automobiles, television sets,

[16] For the United States, reliable annual data on gross domestic product begin in 1869.

telephones, personal computers, and a connection to the Internet. Also, compared with 134 years earlier, education levels were much higher, life expectancy was substantially longer, and a much smaller fraction of the population lived on farms.

We can calculate that the average growth rate of real GDP per person in the United States from 1869 to 2003 was 2.0% per year.[17] This growth rate does not seem all that impressive—it equals the average value from 1960 to 2000 for the 111 countries shown in Figure 3.3. Moreover, the growth rate is much less than the rates of as high as 6% per year achieved by some East Asian countries from 1960 to 2000. Nevertheless, the U.S. growth rate of 2.0%—when sustained for such a long time—was enough to make the United States the world's second richest country.

Living standards in the United States in 2003 would have been much different if the average growth rate of real GDP per person since 1869 had been materially lower or higher than the actual value of 2.0%. To see the consequences of differences in long-run growth rates by 1% per year, imagine first that the average U.S. growth rate had been only 1.0% per year. In this case, the level of real GDP per person in 2003 would have been \$8290, only 3.8 times the value in 1869.[18] In this case, the typical American family would have reasonable food and healthcare but would lack a comfortable home and fine automobile, would be missing an array of pleasant consumer products, and would have lower levels of education.

Suppose, instead, that the U.S. growth rate of real GDP per person since 1869 had been higher by 1% per year and averaged 3.0%. In this case, the level of real GDP per

---

[17] The easiest way to do this is to compute the natural logarithm of the ratio of real GDPs per person, ln(33500/2170) = 2.74, and then divide by the number of years, 134, to get 0.020.

[18] We get this answer by seeing that the natural logarithm of real GDP per person in 2003 equals ln(2170) + 134*0.01 = 7.682 + 1.34 = 9.022. Hence, real GDP per person in 2003 equals the exponential of 9.022 or 8290.

person in 2003 would have been $121,000, 56 times the level in 1869.[19] A real GDP per person of $121,000 may seem out of the realm of imagination—it means that the typical person would have a grand home, a couple nice cars, no problems with expensive private schooling and healthcare, and so on. However, people in 1869 probably would have thought that a real GDP per person of $33,500—the actual value for 2003—was also unimaginable.

Similar calculations apply to other OECD countries, many of which are now nearly as rich as the United States. These countries also came to be rich because their real GDP per person grew for a long time at the unspectacular rate of around 2% per year.

We have considered the growth rate of real GDP per person over a century or more. However, the growth rates were not actually constant over time. To see this, consider the numbers in Table 3.2, which apply to 16 OECD countries, including the United States. The table shows averages for the 16 countries of the growth rates of real GDP per person over 20-year periods, going from 1830 to 1990, although the data are more reliable since 1870. If we start in 1870, we find that the average growth rates are reasonably stable through 1950, ranging between 1.2 and 1.5% per year. However, the average growth rates were much higher in the post-World War II period: 3.7% for 1950-70 and 2.2% for 1970-90.

[19] We compute that the natural logarithm of real GDP per person in 2003 equals ln(2170) + 134*0.03 = 7.682 + 4.02 = 11.702. Then real GDP per person in 2003 is the exponential of 11.702, which equals 121,000.

Table 3.2 **Long-Term Economic Growth in OECD Countries**		
Period	**Growth rate of real GDP per person (percent per year)**	**Number of countries**
1830-1850	0.9	10
1850-1870	1.2	11
1870-1890	1.2	13
1890-1910	1.5	14
1910-1930	1.3	16
1930-1950	1.4	16
1950-1970	3.7	16
1970-1990	2.2	16

Note: The data are from Maddison (1991 and unpublished updates). The 16 countries included are Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany (West), Italy, Japan, Netherlands, Norway, Sweden, Switzerland, United Kingdom, and United States. The numbers are averages for the countries with available data. Fewer countries have data for the earlier periods.

The decline in the growth rate of real GDP per person from 1950-70 to 1970-90 is sometimes called the **productivity slowdown**. However, even the reduced growth rate for 1970-90 is well above the rates experienced before 1950. At present, economists are uncertain whether the growth rate of real GDP per person for the advanced, OECD countries will be heading back down toward the 1.2-1.5% range characteristic of the pre-World War II years or, instead, is likely to stay at 2% per year or more.

In looking at the data, we have observed some important patterns in economic growth. First, some countries, such as in East Asia, grew rapidly from 1960 to 2000 and, thereby, raised their levels of real GDP per person substantially over 40 years. Second, over the same period, other countries—especially in sub-Saharan Africa—grew at low or negative rates and, therefore, ended up with low levels of real GDP per person in 2000. Third, the United States and other OECD countries had high levels of realGDP per person in 2000 mostly because they grew at moderate rates—around 2% per year—for a century or more.

These observations suggest questions that we would like to answer about economic growth.

- What factors caused some countries to grow fast and others to grow slow over periods such as 1960 to 2000? In particular, why did the East Asian countries do so much better than the sub-Saharan African countries?
- How did countries such as the United States and other OECD members sustain growth of real GDP per person of around 2% per year for a century or more?
- What can policymakers do to increase growth rates of real GDP per person?

The answers to these questions could contribute a lot to the living standards of future generations. The theories about economic growth that we turn to next bring us closer to getting these answers.

III. Theory of Economic Growth

Now we build a model of economic growth to help to understand the patterns that we found in the international data. We start by considering the production function, which tells us how goods and services are produced.

A. The production function

We begin our theoretical study of economic growth by considering how a country's technology and factor inputs determine its output of goods and services, measured by its real GDP. This relation is called a **production function.** Formally, this function shows how output or GDP, which we represent by *Y*, depends on the level of the technology and the quantities of factor inputs.

We build a simplified model that has two factor inputs: capital, *K*, and labor, *L*. We think of capital as taking a physical form, such as machines and buildings used by businesses. In an extended setting, we would go further to include **human capital**, which measures the effects of education and training on workers' skills and the effects of medical care, nutrition, and sanitation on workers' health. Since we do not include human capital in the model, the amount of labor input, *L*, is the quantity of work-hours per year for labor of a standard quality and effort. That is, we imagine that, at a point in time, each worker has the same skill. For convenience, we often refer to *L* as the **labor force** or the number of workers—these interpretations are satisfactory if we think of each laborer as working a fixed number of hours per year.

We use the symbol *A* to represent the level of the technology. For given quantities of the factor inputs, *K* and *L*, an increase in *A* raises output. That is, a technologically more advanced economy has a higher level of overall **productivity**. By higher productivity, we mean that output is higher for given quantities of the factor inputs.

Mathematically, we write the production function as

Key equation (production function):

$$Y = A \cdot F(K, L). \tag{3.1}$$

One way to see how output, *Y*, responds to the variables in the production function—the technology level, *A*, and the quantities of capital and labor, *K* and *L*—is to change one of the three variables while holding the other two fixed. Looking at the equation, we see that *Y* is proportional to *A*. Hence, if *A* doubles, while *K* and *L* do not change, *Y* doubles.[20]

For a given technology level, *A*, the function *F(K, L)* determines how additional amounts of capital and labor, *K* and *L*, affect output, *Y*. We assume that each factor is productive at the margin. Hence, *Y* rises with *K* for given *A* and *L*, and *Y* rises with *L* for given *A* and *K*.

The effect on *Y* from a small increase in *K* is called the **marginal product of capital**, which we abbreviate as **MPK**. The MPK tells us how much *Y* rises when *K* increases by one unit, while *A* and *L* do not change. The corresponding effect on *Y* from a small increase in *L* is called the **marginal product of labor** or **MPL**. The MPL tells us how much *Y* rises when *L* increases by one unit, while *A* and *K* do not change. We assume that the two marginal products, MPK and MPL, are greater than zero.

Figure 3.5 shows how output, *Y*, responds to an increase in capital input, *K*. This figure is a graph of the production function, $A \cdot F(K, L)$, from equation (3.1). However, the special feature of this graph is that we are holding constant the values of *A* and *L*. Therefore, the figure shows how increases in *K* affect *Y* when *A* and *L* do not change.

[20] The form in which the technology level, *A*, enters into equation (3.1) is called Hicks neutral, after the British economist John Hicks. In this form, a change in *A* affects the productivity of the two inputs, *K* and *L*, in a parallel or neutral manner.

The curve in Figure 3.5 goes through the origin, because we assume that output is zero if the capital stock is zero. The slope of the curve at any point is the marginal product of capital, that is, the effect on Y from a small increase in K. Since we have assumed that this marginal product, MPK, is always greater than zero, the slope of the curve is positive throughout. We also assume that the slope flattens as the quantity of capital rises. The curve flattens as K rises because we assume that the MPK declines as K rises, for given A and L. This property is known as **diminishing marginal product of capital**.

Figure 3.6 shows the corresponding graph for output, Y, as a function of labor input, L. This figure is again a graph of the production function, $A \cdot F(K, L)$, from equation (3.1). However, the special feature now is that we are holding constant the values of A and K. This figure shows how increases in L affect Y when A and K do not change. Again, the curve goes through the origin, because we assume that output is zero if labor input is zero. The positive slope of the curve at any point is the marginal product of labor, MPL. The flattening of the curve as L rises means that the MPL falls as L increases, for given A and K. Hence, we are assuming **diminishing marginal product of labor**.

Another assumption is that the production function in equation (3.1) exhibits **constant returns to scale** in the two factor inputs, K and L. For example, assume that a business starts with 5 machines, $K = 5$, and 5 workers, $L = 5$. The business has a given technology level, A, and is able to produce an output, Y, of, say, 100 widgets per year. Now, suppose that K and L double, so that the business has $K = 10$ machines and $L = 10$ workers. The technology level, A, is the same as before. Our assumption is that, with twice as many machines and workers and the same technology level, the business can produce twice as much output. That is, output, Y, is now 200 widgets per year.

More generally, if the production function exhibits constant returns to scale, a multiplication of the two factor inputs, K and L, by any positive number leads to a multiplication of output, Y, by the same number. Therefore, if we multiply K and L by the quantity $1/L$ in equation (3.1), we also multiply Y by $1/L$ to get

$$Y/L = A \cdot F(K/L,\ L/L).$$

The value L/L on the right-hand side is constant at the value one and can, therefore, be dropped. The point of writing the production function this way is that it shows that output per worker, Y/L, depends only on the technology level, A, and the quantity of capital per worker, K/L. We can see this property more clearly by defining $y \equiv Y/L$ to be output per worker and $k \equiv K/L$ to be capital per worker and then defining a new function, f, that relates y to k:

$$y = A \cdot f(k). \qquad (3.2)$$

Figure 3.7 shows the graph of output per worker, y, versus capital per worker, k, for a given technology level, A. This graph looks the same as the one in Figure 3.5. The slope of the curve in Figure 3.7 again tells us the effect of more capital on output, that is, it measures the marginal product of capital, MPK. Note that the marginal product of capital diminishes as capital per worker, k, rises.

B. Growth accounting

The production function determines the level of output or GDP, Y, at a point in time for given values of its three determinants: the technology level, A, and the quantities of capital and labor, K and L. However, the production function is also the starting point for our investigation of economic growth. To use the production function to study growth, we use a method called **growth accounting** to consider how growth in Y depends on growth in A, K, and L. Whereas the production function is a relation between the level of Y and the levels of A, K, and L, growth accounting is a relation between the growth rate of Y and the growth rates of A, K, and L.

To begin the analysis of growth accounting, let ΔY represent the change in Y over an interval of time, say, a year. (The symbol Δ, the Greek letter capital delta, represents the change in a variable.) Then the growth rate of Y over the year is given by $\Delta Y/Y$. For example, if $Y = 100$ and $\Delta Y = 1$ over a year, the growth rate is $\Delta Y/Y = 1\%$ per year. Similarly, if we use ΔA, ΔK, and ΔL to represent the changes in technology, capital, and labor, the growth rates of each are $\Delta A/A$, $\Delta K/K$, and $\Delta L/L$, respectively.

Our next task is to explain precisely how $\Delta A/A$, $\Delta K/K$, and $\Delta L/L$ contribute to the growth rate of real GDP, $\Delta Y/Y$. Start with the contribution of technology. We see from the production function,

$$Y = A \cdot F(K, L), \tag{3.1}$$

that Y would grow at the same rate as A if K and L were constant. For example, if $\Delta A/A = 1\%$ per year and K and L are constant, then $\Delta Y/Y = 1\%$ per year. Even if K and L are changing, equation (3.1) tells us that a higher growth rate of A would contribute to a higher growth rate of Y. In particular, if $\Delta A/A$ is higher by 1% per year, then $\Delta Y/Y$ is higher by 1% per year, for given growth rates of capital and labor, $\Delta K/K$ and $\Delta L/L$.

Now consider the contributions to the growth of real GDP from growth in capital and labor. We know that $\Delta Y/Y$ depends positively on $\Delta K/K$ and $\Delta L/L$. To be more precise, suppose that the contribution of growth in capital to the growth of real GDP is given by $\alpha \cdot \Delta K/K$ where α (the Greek letter alpha) is greater than zero. Under reasonable conditions detailed in the appendix to this chapter, α equals the share of capital income in the economy's total income. Therefore, α is between zero and one. For example, if $\alpha=1/3$—a commonly assumed value for the share of capital income—then a growth rate of capital, $\Delta K/K$, of 1% a year would contribute (1/3)% per year to the growth rate of real GDP, $\Delta Y/Y$.

Similarly, the contribution of labor to the growth of real GDP is $\beta \cdot (\Delta L/L)$, where β (the Greek letter beta) is greater than zero. Under the conditions explored in the appendix to this chapter, β equals the share of labor income in the economy's total income. Therefore, β is between zero and one. Moreover, since all income goes to capital or labor, the two income shares must add to one:

$$\text{share of capital income} + \text{share of labor income} = 1$$

$$\alpha + \beta = 1.$$

As an example, if $\alpha=1/3$, then $\beta=2/3$, which is a commonly assumed value for the share of labor income. In this case, a growth rate of labor, $\Delta L/L$, of 1% a year would contribute (2/3)% per year to $\Delta Y/Y$.

We now have all the pieces necessary to write down the growth-accounting equation:

$$\Delta Y/Y = \Delta A/A + \alpha \cdot (\Delta K/K) + \beta \cdot (\Delta L/L).$$

That is, the growth rate of real GDP, $\Delta Y/Y$, equals the growth rate of technology, $\Delta A/A$, plus the contributions from the growth of capital, $\alpha \cdot (\Delta K/K)$, and the growth of labor,

$\beta \cdot (\Delta L/L)$. If we rearrange the condition $\alpha + \beta = 1$ to substitute $1-\alpha$ for β on the right-hand side of the equation, we get

Key equation (growth accounting):

$$\Delta Y/Y = \Delta A/A + \alpha \cdot (\Delta K/K) + (1-\alpha) \cdot (\Delta L/L). \tag{3.3}$$

Equation (3.3) says that we can break down $\Delta Y/Y$ into $\Delta A/A$ and a weighted average of the growth rates of capital and labor, $\alpha \cdot \Delta K/K$ and $(1-\alpha) \cdot \Delta L/L$. The growth rate of capital gets the weight α (which is capital's share of income), and the growth rate of labor gets the weight $1-\alpha$ (which is labor's share of income).

We now simplify the analysis by assuming that the coefficient α, which equals capital's share of income, is constant. That is, we assume that this share does not vary as the economy grows. This assumption implies that the share going to labor, which equals $1-\alpha$, is also constant. The constancy of income shares does not always hold in the real world, but it does work as a reasonable approximation for the United States and many other countries. In the appendix to this chapter, we show that the constancy of income shares applies for a commonly assumed form of the production function, $A \cdot F(K, L)$.

C. The Solow growth model

We learned from growth accounting in equation (3.3) that the growth rate of real GDP, $\Delta Y/Y$, depends on the growth rate of technology, $\Delta A/A$, and the growth rates of capital and labor, $\Delta K/K$ and $\Delta L/L$. To go from growth accounting to a theory of economic growth, we have to explain the growth of technology, capital, and labor. We begin this explanation by constructing the **Solow growth model**. The first version of this model emphasizes the growth rate of capital, $\Delta K/K$. Later we bring in growth of labor, $\Delta L/L$, and technology, $\Delta A/A$.

Intellectual history of the Solow growth model

The Solow model was created during the 1950s by the MIT economist Robert Solow. This research led eventually to a Nobel Prize in 1987 for "contributions to the theory of economic growth." The Solow model was extended during the 1960s, especially by Tjalling Koopmans and David Cass, and became known as the **neoclassical growth model**. The Solow model and the 1960's extensions were actually anticipated in theoretical work done by the mathematician Frank Ramsey in the 1920s. Hence, this growth model is often called the **Ramsey model**. Unfortunately, Ramsey's brilliant research was cut short by his early death in 1930 at age 26.

1. Setup of the model. The Solow model makes several simplifying assumptions. First, the labor input, L, equals the **labor force**, which is the number of persons who are seeking work. That is, the model does not allow for unemployment—labor input equals the labor force, all of which is employed.[21]

The relation between the labor force, L, and the population is given by

$$labor\ force,\ L = (labor\ force/population) \cdot population.$$

The expression (labor force/population) is the **labor-force participation rate**. In the United States in recent years, the labor-force participation rate has been close to one-half.[22] The second assumption in the Solow model is that this participation rate is constant. In this case, the above equation tells us that the growth rate of labor input, L, equals the growth rate of the population.

Third, the model ignores a role for government, so that there are no taxes, public expenditures, government debt, or money. Fourth, the model assumes that the economy is closed, so that there is no international trade in goods and services or in financial assets.

To begin our analysis of the Solow model, consider the growth-accounting equation:

$$\Delta Y/Y = \Delta A/A + \alpha\cdot(\Delta K/K) + (1\text{-}\alpha)\cdot(\Delta L/L). \quad (3.3)$$

As already mentioned, we start by focusing on growth in capital, $\Delta K/K$. Therefore, we assume that the technology level is constant, so that $\Delta A/A = 0$. We also simplify initially by ignoring population growth, so that labor input is constant and, hence, $\Delta L/L = 0$. Given these assumptions, the growth-accounting equation simplifies to

$$\Delta Y/Y = \alpha\cdot(\Delta K/K). \quad (3.4)$$

---

[21] The important assumption is that the unemployment rate is constant, not necessarily zero. For example, if 96% of the labor force is always employed, labor input would always be a constant multiple of the labor force and would grow at the same rate as the labor force.

[22] For example, in 2003, the sum of civilian employment (137.7 million) and the military (1.4 million) was 48% of the total population (including military overseas) of 291.1 million.

Hence, in this first version of the Solow model, economic growth arises only from increases in the stock of capital.

We see from equation (3.4) that, to determine the growth rate of real GDP, $\Delta Y/Y$, we have to determine the change in the stock of capital, ΔK. This change will depend on the economy's **saving**, which is the income that is not consumed. Hence, we now consider how saving is determined.

In a later analysis (in chapter 7), we work out saving behavior from micro foundations that start with the optimizing choices of individual households. However, for now, we simplify by using Solow's assumption that households divide up their income in a fixed proportion s to saving and 1-s to consumption, C. For the economy as a whole, the real GDP, Y, corresponds to total real income. Some of this income goes to workers and some to owners of capital. However, all of the income must flow eventually to households, partly in their role as workers and partly in their role as owners of capital (or owners of businesses). Hence, the total of household real income is Y, and—if the saving rate is s for all forms of income—the aggregate amount saved is sY.

The economy's total real income, Y, equals consumption, C, plus saving, sY. In a closed economy with no government sector, the real GDP, Y, must also be either consumed or invested. That is, the goods and services produced are used for only two purposes: consumption and outlays on capital goods, which are **gross investment**, I. Therefore, we have

$$Y = C + sY \text{ (income is consumed or saved) and}$$

$$Y = C + I \text{ (output is consumed or invested).}$$

Consequently,

$$C + sY = C + I.$$

If we cancel out the variable C on the two sides of the equation, we get an equality between saving and gross investment:

(3.5) $$sY = I.$$

Gross investment, I, would equal the change in the capital stock, ΔK, if there were no **depreciation** or wearing out of the existing capital. More realistically, capital stocks depreciate over time. We use the symbol δ (the Greek letter delta) to represent the rate of depreciation. The depreciation rate is a positive number, so that $\delta > 0$, and δK is the amount of capital that depreciates each year. The value of δ depends on the type of building or machine, but a reasonable average number is 5% per year.

The change in the capital stock is given by

change in capital stock = gross investment minus depreciation,

so that

$$\Delta K = I - \delta K.$$

The quantity $I - \delta K$—gross investment less depreciation—is **net investment**. Substituting $I = sY$ from equation (3.5), we get

(3.6) $$\Delta K = sY - \delta K.$$

If we divide through each side of the equation by K, we get a formula for the growth rate of the capital stock:

Key equation (Solow growth model):

(3.7) $$\Delta K/K = s \cdot (Y/K) - \delta.$$

Equation (3.7) is the key relation in the Solow growth model. Because of the importance of this equation, we should take a moment to be sure that we understand the various terms. On the left-hand side, the numerator, ΔK, is the change in the capital stock over a year. Therefore, this variable is a flow that has units of goods per year. The denominator, K, is a stock—the stock of capital—and has units of goods. Hence, $\Delta K/K$ has units of

(goods per year)/goods = per year.

For example, a value for $\Delta K/K$ of 0.02 per year means that the capital stock is growing at 2% per year.

The two terms on the right-hand side of equation (3.7) are each determinants of the growth rate of capital. Hence, each of these terms must also have units of per year. These units are clear for the depreciation rate, δ, which we mentioned might be 0.05 (or 5%) per year.

Consider now the term $s \cdot (Y/K)$, which is the product of the saving rate, s, and Y/K. The saving rate, s, is a pure number, for example, 0.2 if households save 20% of their income. The term Y/K—output per unit of capital—is called the **average product of capital**. The units for Y—a flow variable—are goods per year and those for K—a stock variable—are goods. Therefore, the average product of capital has units of

(goods per year)/goods = per year.

Since s is a pure number, the units of $s \cdot (Y/K)$ are the same as the units of Y/K, that is, per year, just like $\Delta K/K$.

In much of the analysis, we will find it useful to express the average product of capital, Y/K, in terms of real GDP per worker, y, and capital per worker, k. The relation is

$$Y/K = \frac{Y/L}{K/L}$$

(3.8) $$Y/K = \frac{y}{k}.$$

If we substitute this result for Y/K into equation (3.7), we get

Key equation (Solow growth model):

(3.9) $$\Delta K/K = s\cdot(y/k) - \delta.$$

The growth rate of the capital stock, $\Delta K/K$, determines the growth rate of real GDP, $\Delta Y/Y$, from the growth-accounting equation:

(3.4) $$\Delta Y/Y = \alpha\cdot(\Delta K/K).$$

Therefore, if we substitute for $\Delta K/K$ from equation (3.9), we get

(3.10) $$\Delta Y/Y = \alpha\cdot[s\cdot(y/k) - \delta].$$

2. The transition and the steady state. To recapitulate, equation (3.10) says that the growth rate of real GDP, $\Delta Y/Y$, depends on the saving rate, s, the depreciation rate, δ, the capital share of income, α, and the average product of capital, y/k. We have assumed that s, δ, and α are constants. Therefore, the only reason that $\Delta Y/Y$ varies over time is that the average product of capital, y/k, varies. We now consider how this average product depends on the capital per worker, k. In this way, we shall find that changes over time in k lead to changes in y/k and, thereby, to changes in $\Delta Y/Y$.

We considered before the marginal product of capital, MPK, which is the ratio of a change in real GDP, ΔY, to a change in capital, ΔK. Geometrically, the marginal product was given by the slope of the production function, which is shown in Figure 3.7

and reproduced in Figure 3.8. In this new graph, we compute the average product of capital, y/k, as the ratio of y (the variable on the vertical axis) to k (the variable on the horizontal axis). This ratio equals the slope of a straight line from the origin to the production function. The graph shows two such lines, corresponding to capital per worker of $k_1 = K_1/L$ and $k_2 = K_2/L$, where k_2 is greater than k_1. (Recall that labor input is constant at L.) The graph shows that the average product of capital, y/k, declines as capital per worker, k, rises. This **diminishing average product of capital** is analogous to the diminishing marginal product, which we discussed before.

We can show diagrammatically how the growth rate of capital, $\Delta K/K$, is determined in equation (3.9) by graphing each of the two terms on the right-hand side of the equation versus the capital per worker, k. In the first term, $s\cdot(y/k)$, the crucial property is the one that we just derived: the average product of capital, y/k, diminishes as k rises. Hence, the curve for $s\cdot(y/k)$ is downward sloping versus k, as shown in Figure 3.9. The second term on the right-hand side of equation (3.9), δ, appears in the graph as a horizontal line.

To study how the growth rate of capital, $\Delta K/K$, changes over time we have to know the level of capital that the economy has at some initial date, which we label as date 0. That is, the economy starts with an accumulated stock of capital in the forms of machines and buildings. We represent this starting stock by $K(0)$. Hence, in Figure 3.9, the initial capital per worker is $k(0) = K(0)/L$. The corresponding initial level of real GDP per worker is given from equation (3.2) by $y(0) = Y(0)/L = A\cdot f[k(0)]$.

In Figure 3.9, the growth rate of capital, $\Delta K/K$, is the vertical distance between the $s\cdot(y/k)$ curve and the horizontal line, δ. (See equation [3.9].) We assume that, when

$k = k(0)$, the curve lies above the line. In this case, the capital stock is growing initially. That is, $\Delta K/K$ is greater than zero and is given by the distance marked by the green arrows. This positive growth rate means that the stock of capital, K, increases over time.

Since labor input, L, is constant, capital per worker, $k = K/L$, grows at the same rate as K—that is, $\Delta k/k = \Delta K/K$. Hence, k rises over time and moves rightward in Figure 3.9. Notice that the distance between the curve and the horizontal line diminishes over time. Since this distance equals the growth rate, $\Delta K/K = \Delta k/k$, we have shown that the growth rate slows down over time. This result is an important property of the Solow model.

Eventually, the increase in capital per worker, k, eliminates the gap between the $s\cdot(y/k)$ curve and the δ line. The gap becomes zero when k equals the value k^* shown in Figure 3.9.[23] At this point, the growth rate, $\Delta K/K = \Delta k/k$, equals zero. Therefore, k no longer moves to the right—because with $\Delta k/k = 0$, k stays forever at the value k^*. For this reason, we call k^* the **steady-state capital per worker**.

The results tell us that capital per worker, k, will follow a **transition path** from its initial value, $k(0)$, to its steady-state value, k^*. Figure 3.10 shows this transition path as the red curve. Note that k starts at $k(0)$, rises over time, and eventually gets close to k^*, which is shown as the dashed blue line.

Since the quantity of labor, L, is constant, the capital stock, K, is also constant in the steady state. This steady-state capital stock, K^*, is determined from the relation $k^* = K^*/L$ to be $K^* = Lk^*$.

Recall that the formula for $\Delta K/K$ is

[23] More precisely, k would get very close to k^* but not actually reach it in finite time. A more accurate statement is that k gradually approaches k^*.

(3.7) $$\Delta K/K = s\cdot(Y/K) - \delta.$$

In the steady state, $\Delta K/K$ equals zero. Therefore, the right-hand side of equation (3.7) equals zero in the steady state:

$$s\cdot(Y^*/K^*) - \delta = 0.$$

If we move δ to the right-hand side and multiply through by K^*, we get

(3.11) $$sY^* = \delta K^*$$

steady-state saving = steady-state depreciation.

We know that saving always equals gross investment:

(3.5) $$sY = I.$$

Therefore, we can substitute $I^* = sY^*$ in equation (3.11) to get

(3.12) $$I^* = \delta K^*$$

steady-state gross investment = steady-state depreciation.

In other words, steady-state gross investment, I^*, is just enough to cover the steady-state depreciation of capital, δK^*. Recall that the change in the capital stock equals net investment:

$$\Delta K = I - \delta K.$$

Therefore, $I^* = \delta K^*$ implies, as we already know, that the change in the capital stock in the steady state, $(\Delta K)^*$, is zero.

Our analysis allows us to think of the process of economic growth in the Solow model as having two phases. In the first phase, there is a transition from an initial capital per worker, $k(0)$, to its steady-state value, k^*. This transition is shown by the red curve in Figure 3.10. During this transition, the growth rate, $\Delta K/K = \Delta k/k$, is greater than zero but declines gradually toward zero. Since K is rising, net investment is greater than zero, that is, gross investment, I, exceeds depreciation, δK. In the second phase, the economy is in (or close to) the steady state, represented by the dashed blue line in Figure 3.10. In this phase, $\Delta K/K = \Delta k/k = 0$, net investment is zero, and gross investment, I^*, just covers depreciation, δK^*.

Our goal was to determine how the growth rate of real GDP, $\Delta Y/Y$, varies over time. We can now reach this goal because, from equation (3.4), $\Delta Y/Y$ equals the growth rate of capital, $\Delta K/K$, multiplied by the constant α (where $0<\alpha<1$):

$$\Delta Y/Y = \alpha\cdot(\Delta K/K). \tag{3.4}$$

Therefore, everything that we said about $\Delta K/K$ applies also to $\Delta Y/Y$, once we multiply by α. In particular, starting at the initial capital per worker $k(0)$ shown in Figure 3.9, we have that $\Delta Y/Y$ starts out positive, then declines as the stock of capital, K, and the real GDP, Y, rise. Eventually, when capital per worker, k, reaches its steady-state value, k^*, $\Delta Y/Y$ falls to zero (because equation [3.4] implies $\Delta Y/Y = 0$ when $\Delta K/K = 0$).

When capital per worker, k, is constant at its steady-state value, k^*, real GDP per worker, y, is constant at its steady-state value, which we represent by y^*. The value of y^* is determined from equation (3.2) as $y^* = A\cdot f(k^*)$. Since the quantity of labor, L, is constant, the real GDP, Y, is also constant in the steady state. This steady-state level, which we represent by Y^*, is determined from the relation $y^* = Y^*/L$ to be $Y^* = L\cdot y^*$.

In terms of Figure 3.10, we can view the transition as applying to real GDP per worker, y, as well as to capital per worker, k. That is, the red curve also describes the transition from the initial real GDP per worker, $y(0)$, to its steady-state value, y^*.

Using algebra

We have seen how Figure 3.9 determines the steady-state capital per worker, k^*. We can also determine k^* algebraically. If we set $\Delta K/K = 0$ in equation (3.9), the right-hand side of the equation must be zero, so that $s\cdot(y/k) - \delta = 0$. If we rearrange the terms in this equation and divide by s, we find that the average product of capital in the steady state is

(3.13) $$y^*/k^* = \delta/s.$$

If we use equation (3.2) to get $y^* = A\cdot f(k^*)$ and substitute $A\cdot f(k^*)$ for y^*, we get that the steady-state capital per worker, k^*, must satisfy the equation

(3.14) $$A\cdot f(k^*)/k^* = \delta/s.$$

This algebraic result for k^* will be helpful in the next chapter when we work further with the Solow model.

D. Summing up

We motivated our study of the theory of economic growth from observations about the importance of growth for standards of living. Now we have constructed the Solow growth model and are ready to work with it to understand how economic variables influence growth. We begin to put the model to use in the next chapter.

Questions and Problems

Mainly for review

3.1. What is a production function? In what way does it represent a relation between factors inputs and the level of output?

3.2. Explain the concepts of marginal and average product of capital. What is the difference between the two? Is the average product always greater than the marginal product?

3.3. Does a positive saving rate, $s > 0$, mean that the capital stock, K, rises over time? Explain by referring to equation (3.6).

3.4. Does a positive saving rate, $s > 0$, mean that output, Y, and the capital stock, K, grow in the long run? Explain.

3.5. Explain why an increase in capital per worker, k, reduces the growth rate of the capital stock, $\Delta K/K$. How does this result depend on diminishing productivity of capital?

Problems for discussion

3.x. Constant returns to scale

We have assumed that the production function, $A \cdot F(K, L)$, exhibits constant returns to scale. That is, if we multiply the inputs, K and L, by any positive number, we multiply output, Y, by the same number. Show that this condition implies that we can write the production function as in equation (3.2):

$$y = A \cdot f(k),$$

where $y = Y/L$ and $k = K/L$.

3.x. Determination of steady-state capital per worker

Consider the steady-state capital per worker, k^*, determined in Figure 3.9. How is k^* affected by the following:

a. An increase in the saving rate, s.

b. An increase in the technology level, A.

c. An increase in the depreciation rate, δ.

3.x. Growth with a Cobb-Douglas production function

Suppose that the production function takes the form, $Y = A \cdot F(K,L) = AK^{\alpha}L^{1-\alpha}$, where $0 < \alpha < 1$. This form is the Cobb-Douglas production function, discussed in the appendix to this chapter.

a. In the steady state, $\Delta K/K$, given by equation (3.9), equals zero. Use this condition, along with the form of the production function, to get a formula for the steady-state capital and output per worker, k^* and y^*.

b. What are the steady-state values of investment, I^*, and consumption, C^*?

c. Use equation (3.9) to work out a formula for the growth rate of the capital stock, $\Delta K/K$. Can you show that $\Delta K/K$ falls during the transition as k rises? What happens during the transition to the growth rate of output, $\Delta Y/Y$?

3.x. Growth without diminishing productivity of capital

Suppose that the production function is $Y = AK$ (the so-called AK model).

a. What is the condition for the growth rate of capital, $\Delta K/K$, in equation (3.9)? What does the $s\cdot(y/k)$ curve look like in Figure 3.9?

b. What are the growth rates of capital and output, $\Delta K/K$ and $\Delta Y/Y$? Are these growth rates greater than zero? Do these growth rates decline during a transition?

c. Discuss how your results relate to diminishing productivity of capital. Is it plausible that diminishing productivity would not apply?

Appendix

A. The growth-accounting equation

The growth-accounting equation is

(3.3) $$\Delta Y/Y = \Delta A/A + \alpha\cdot(\Delta K/K) + (1\text{-}\alpha)\cdot(\Delta L/L).$$

We derive this equation more formally here and work out a formula for the coefficient α.

The production function is

(3.1) $$Y = A\cdot F(K, L).$$

The form of this equation tells us that, for given K and L, an increase in the growth rate of technology, $\Delta A/A$, by 1% per year raises the growth rate of real GDP, $\Delta Y/Y$, by 1% per year. This reasoning explains why the term $\Delta A/A$ appears in equation (3.3).

Consider now the effect from changes in K, when A and L are fixed. If K increases by the amount ΔK, while A and L do not change, the increase in real GDP equals ΔK multiplied by the marginal product of capital, MPK:

$$\Delta Y = MPK\cdot\Delta K.$$

To get the growth rate of Y, we have to divide each side by Y:

$$\Delta Y/Y = (MPK/Y)\cdot\Delta K.$$

Then, if we multiply and divide by K on the right-hand side, we get

$$\Delta Y / Y = (\frac{MPK \bullet K}{Y}) \bullet (\Delta K / K).$$

This result determines $\Delta Y/Y$, when K is growing but A and L are constant. More generally, it tells us the contribution of $\Delta K/K$ to $\Delta Y/Y$, even when A and L are changing. That is, to get the contribution of $\Delta K/K$ to $\Delta Y/Y$, we have to multiply $\Delta K/K$ by the term $(MPK\cdot K)/Y$. Therefore, in equation (3.3), we must have that

(3.15) $$\alpha = (MPK\cdot K)/Y.$$

In a competitive economy, capital's marginal product, MPK, would be equated to the real rental price per unit of capital. (We work out this result in chapter 6.) In that case, the term $MPK \cdot K$ equals the total payments to capital, and α equals the share of these payments to capital in the economy's total real income, Y. That is,

$$\alpha = (MPK \cdot K)/Y$$

$$= (\text{payments to capital})/(\text{total income})$$

$$= \text{share of capital in total income}.$$

Since this share is a positive fraction, we have that $0<\alpha<1$.

Now we consider the contribution to the growth of real GDP from growth in labor input. If L increases by the amount ΔL, while A and K are fixed, the increase in real GDP equals ΔL multiplied by the marginal product of labor:

$$\Delta Y = MPL \cdot \Delta L.$$

If we divide through by Y, we get

$$\Delta Y/Y = (MPL/Y) \cdot \Delta L.$$

Then, if we multiply and divide by L on the right-hand side, we get

$$\Delta Y / Y = \left(\frac{MPL \bullet L}{Y}\right) \bullet (\Delta L / L).$$

Therefore, to get the contribution of $\Delta L/L$ to $\Delta Y/Y$, we have to multiply $\Delta L/L$ by the term $(MPL \cdot L)/Y$. Hence, in equation (3.3), we must have that

$$1 - \alpha = (MPL \cdot L)/Y.$$

In a competitive economy, labor's marginal product, MPL, would be equated to the real wage rate. (We work out this result in chapter 6.) Therefore, the term $MPL \cdot L$ equals the total payments to labor, and $(MPL \cdot L)/Y$ is the share of these labor payments in the economy's total real income, Y:

$$(MPL \cdot L)/Y = (payments\ to\ labor)/(total\ income)$$

$$= share\ of\ labor\ in\ total\ income.$$

Since we already know that α equals the share of capital in total income, we get from

$$1 - \alpha = (MPL \cdot L)/Y$$

that

$$1 - share\ of\ capital\ in\ total\ income = share\ of\ labor\ in\ total\ income.$$

If we move the share of capital in total income to the right-hand side, we get that the two income shares add to one. This result is correct because all of the real income in the economy has to go to either capital or labor.[24]

B. The Solow residual

In the economics literature, the term $\Delta A/A$ in equation (3.3) is often called **total factor productivity growth** or **TFP growth**. This concept comes from Solow (1957) and is also often called the **Solow residual**. This terminology arises because a rearrangement of equation (3.3) leads to

$$\Delta A/A = \Delta Y/Y - \alpha \cdot (\Delta K/K) - (1-\alpha) \cdot (\Delta L/L). \tag{3.16}$$

Hence, $\Delta A/A$ can be computed as the residual after we take the growth rate of real GDP, $\Delta Y/Y$, and subtract out the contributions to growth from the changing factor inputs, $\alpha \cdot (\Delta K/K) + (1-\alpha) \cdot (\Delta L/L)$. Economists have calculated these Solow residuals for various countries and time periods.

[24] More formally, it is possible to show from calculus that, if capital and labor are each paid their marginal products and the production function exhibits constant returns to scale (as we have assumed), then the real income payments to capital and labor will equal the economy's output.

C. The Cobb-Douglas production function

We assumed in our analysis of the Solow model that capital's share of income, α, was constant. That is, this share did not change as capital per worker, k, varied. We can show that this assumption corresponds to a particular form of the production function:

$$Y = A \cdot F(K, L)$$

$$Y = AK^{\alpha}L^{1-\alpha}. \tag{3.17}$$

In this form, the constant α appears as the exponent on capital, K, whereas $1-\alpha$ appears as the exponent on labor, L. We assume that α is a fraction, so that $0<\alpha<1$. This form of the production function is assumed by economists in many theoretical and empirical studies.

The form in equation (3.17) is called the **Cobb-Douglas production function**, named after the economist and U.S. Senator Paul Douglas, who allegedly teamed up with a mathematician named Cobb. It is easy to show that the Cobb-Douglas production function satisfies constant returns to scale. (Multiply K and L each by two and see what happens to Y.) In terms of real GDP and capital per worker, y and k, the Cobb-Douglas production function is

$$\begin{aligned} y &= Y/L \\ &= AK^{\alpha}L^{1-\alpha} \cdot (1/L) \\ &= AK^{\alpha}L^{1-\alpha} \cdot L^{-1} \\ &= AK^{\alpha}L^{-\alpha} \\ &= A \cdot (K/L)^{\alpha} \end{aligned}$$

$$y = Ak^{\alpha}. \tag{3.18}$$

We can show using calculus that the exponent α that appears in the Cobb-Douglas production function in equation (3.17) satisfies equation (3.15):

(3.15) $$\alpha = (MPK \cdot K)/Y.$$

To verify this result, recall that MPK is the effect on Y from a change in K, for given values of A and L. If we take the derivative of $Y = AK^{\alpha}L^{1-\alpha}$ with respect to K, while holding fixed A and L, we get

$$\begin{aligned} MPK &= dY/dK \\ &= \alpha AK^{\alpha-1}L^{1-\alpha} \\ &= \alpha AK^{\alpha}K^{-1}L^{1-\alpha} \\ &= \alpha AK^{\alpha}L^{1-\alpha} \cdot (1/K) \\ &= \alpha \cdot (Y/K). \end{aligned}$$

Therefore, we have

$$\begin{aligned} (MPK \cdot K)/Y &= [\alpha \cdot (Y/K) \cdot K]/Y \\ &= \alpha, \end{aligned}$$

as in equation (3.15).

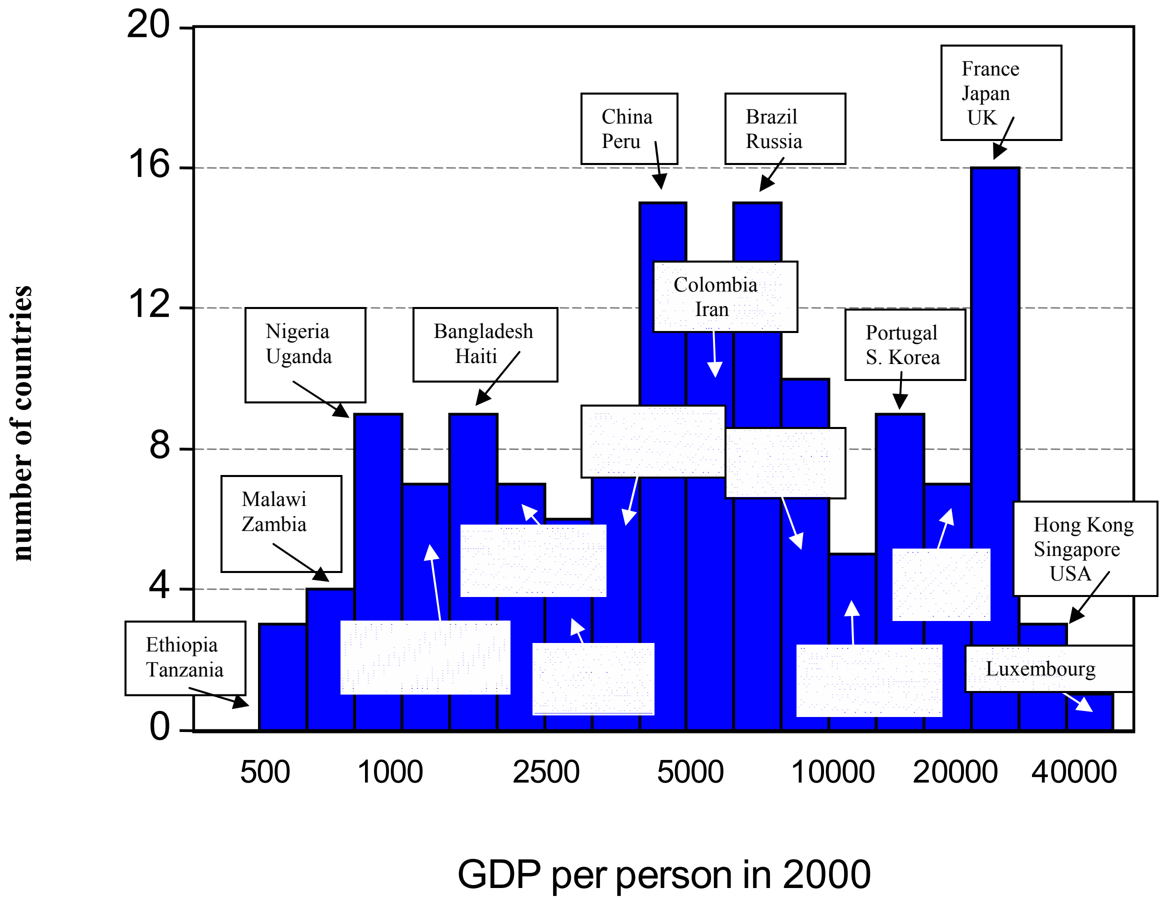


Figure 3.1

The World's Distribution of Real GDP per Person in 2000

The graph shows the distribution of real gross domestic product (GDP) per person for 147 countries in 2000. The horizontal axis is in 1995 U.S. dollars and uses a logarithmic scale. Representative countries are indicated for the various ranges of real GDP per person.

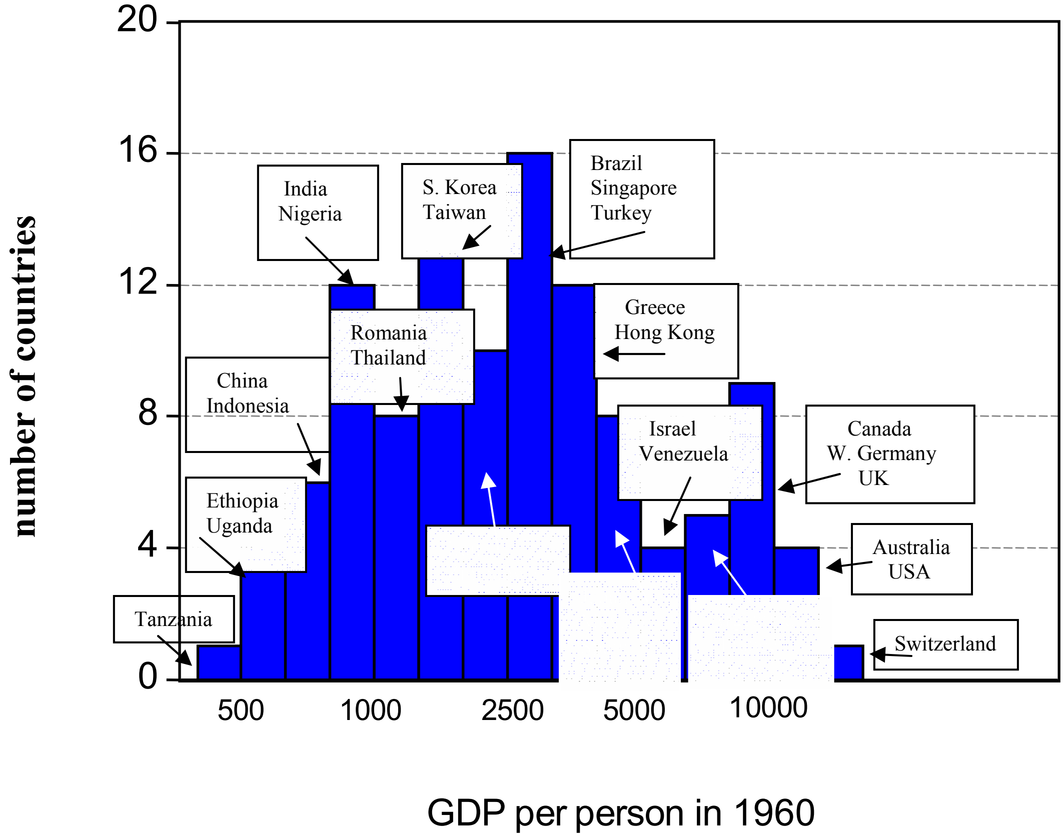


Figure 3.2

The World's Distribution of Real GDP per Person in 1960

The graph shows the distribution of real GDP per person for 113 countries in 1960. The horizontal axis is in 1995 U.S. dollars and uses a logarithmic scale. Representative countries are indicated for the various ranges of real GDP per person.

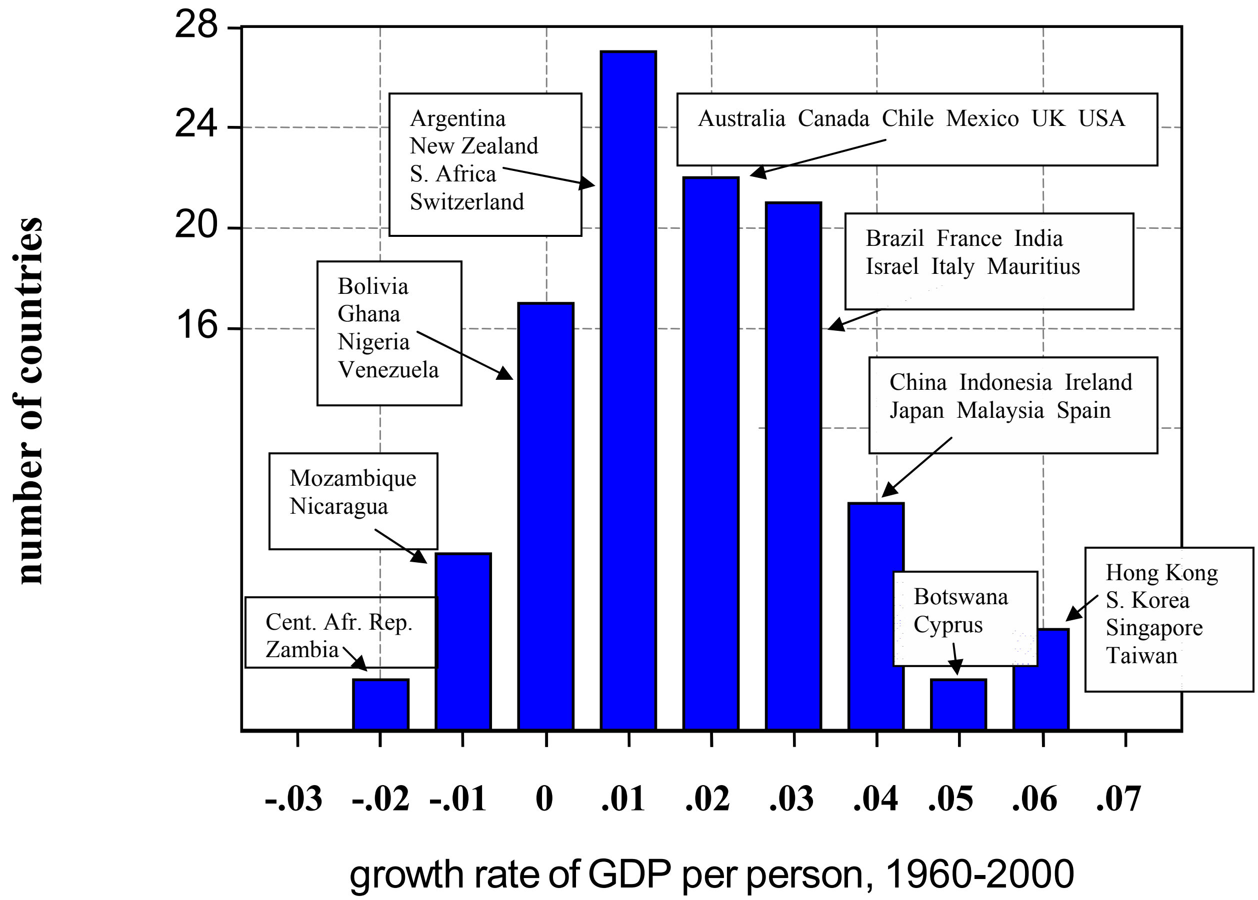


Figure 3.3

The World's Distribution of Growth Rates of Real GDP per Person, 1960-2000

The graph shows the distribution of the annual growth rate of real GDP per person for 111 countries from 1960 to 2000. Representative countries are indicated for the various ranges of growth rates.

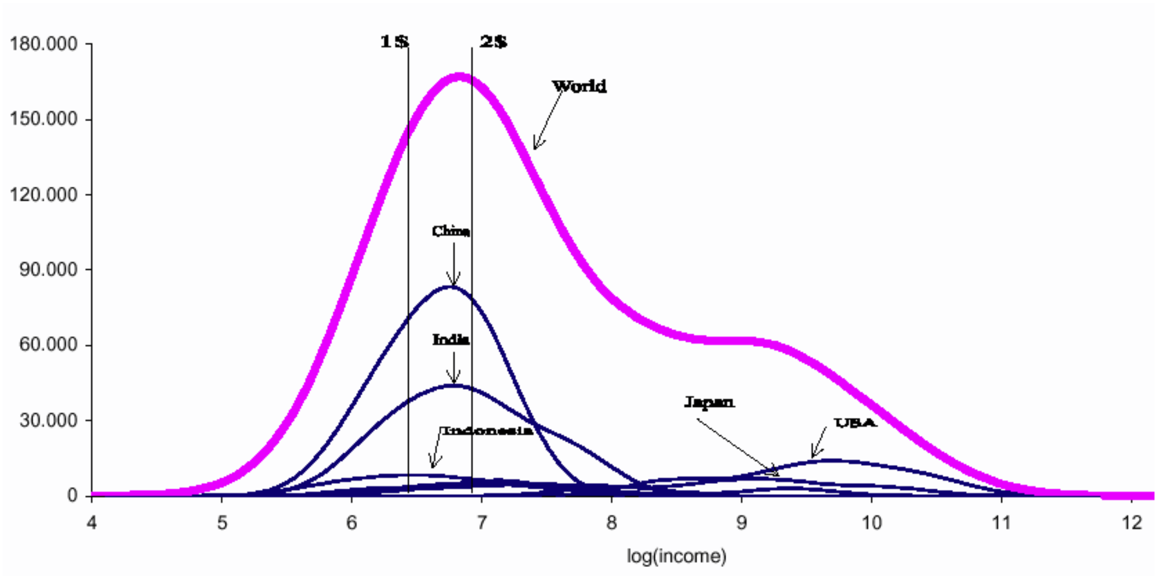


Figure 3.4a
World Income Distribution in 1970

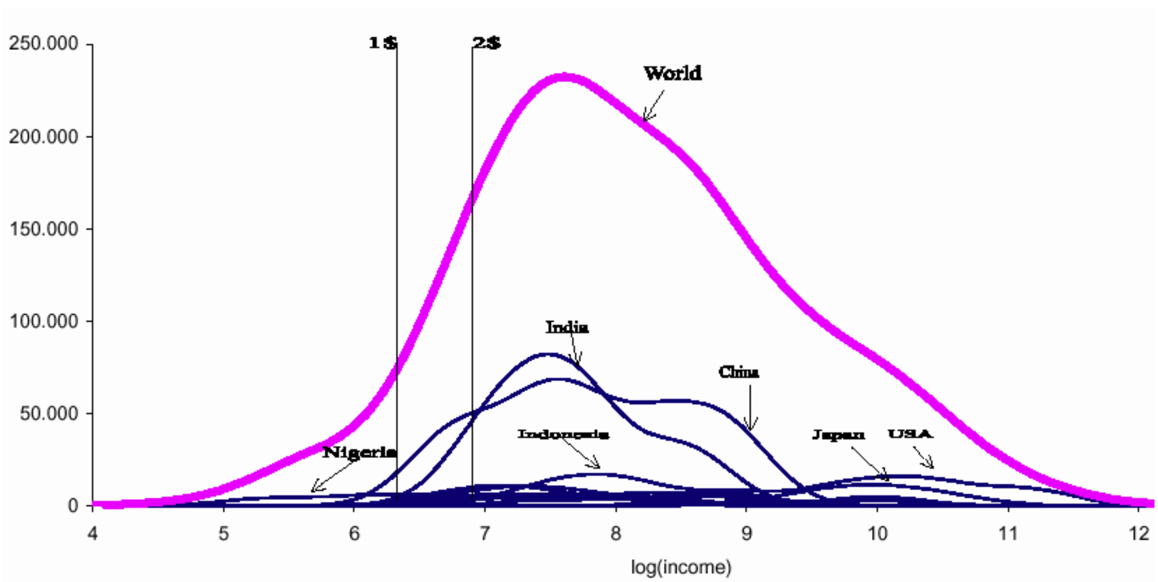


Figure 3.4b
World Income Distribution in 1998

Figure 3.4a is for 1970 and Figure 3.4b is for 1998. These graphs come from Sala-i-Martin (2002). In each case, the horizontal axis plots the log of real income in 1985 U.S. dollars. For the upper curves in the two figures, the vertical axis shows the number of people in the world who had each level of income. The vertical lines marked \$1 and \$2 show the annual real incomes that correspond to the standard poverty lines of \$1 and \$2 per day. The income distributions for a few large countries are shown separately. (Nigeria is hidden beneath the Asian countries in Figure 3.4a.) The values shown on the upper curves for numbers of people in the world are the vertical sums of the numbers of people in all of the individual countries. However, only a few of the countries can be discerned on these graphs.

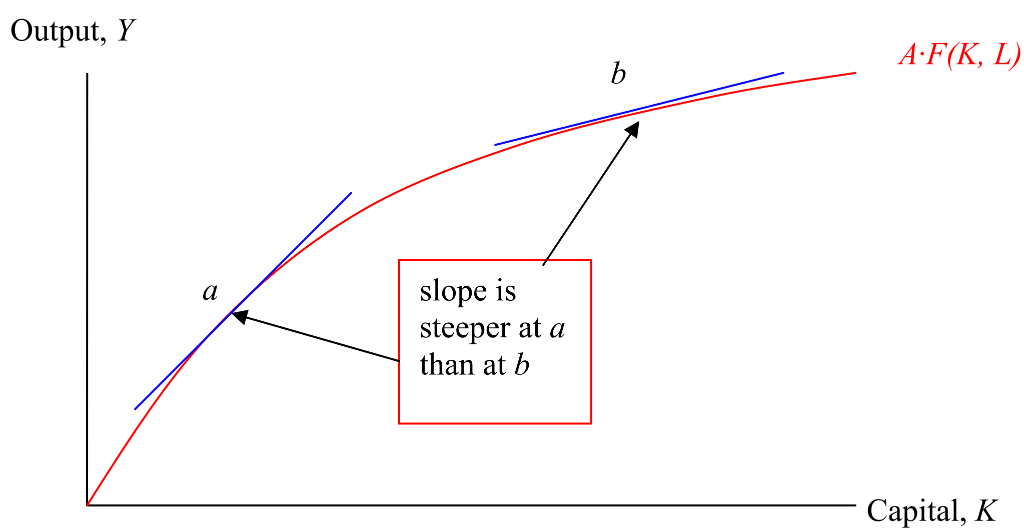


Figure 3.5

The Production Function in Terms of Capital Input

The curve shows the effect of capital input, *K*, on output, *Y*. The technology level, *A*, and the quantity of labor input, *L*, are held fixed. Therefore, the slope of the curve at any point is the marginal product of capital, MPK. This slope gets less steep as *K* rises because of diminishing marginal product of capital. Therefore, the slope at point *a* is greater than that at point *b*.

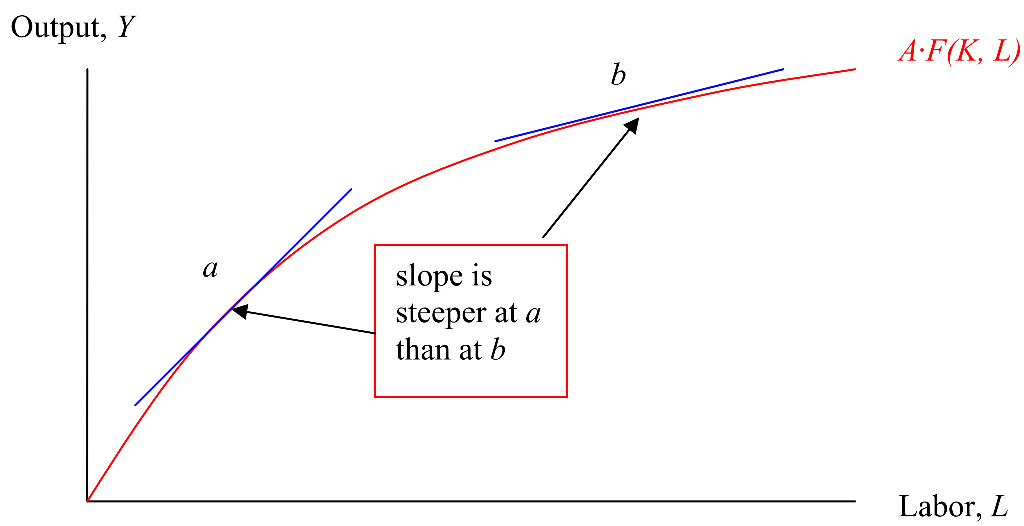


Figure 3.6

The Production Function in Terms of Labor Input

The curve shows the effect of labor input, L, on output, Y. The technology level, A, and the quantity of capital input, K, are held fixed. Therefore, the slope of the curve at any point is the marginal product of labor, MPL. This slope gets less steep as L rises because of diminishing marginal product of labor. Therefore, the slope at point *a* is greater than that at point *b*.

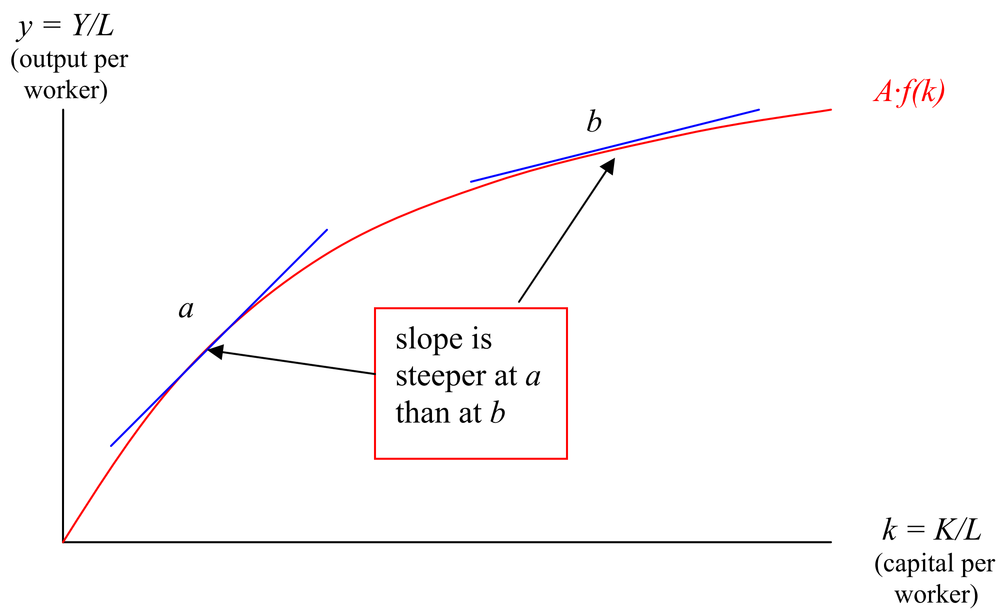


Figure 3.7

Output per Worker versus Capital per Worker

This method for showing the production function plots output per worker, $y = Y/L$, against capital per worker, $k = K/L$. The technology level, A, is held fixed. The slope of the curve at any point is the marginal product of capital, MPK. This slope gets less steep as k rises because of diminishing marginal product of capital. Therefore, the slope at point a is greater than that at point b.

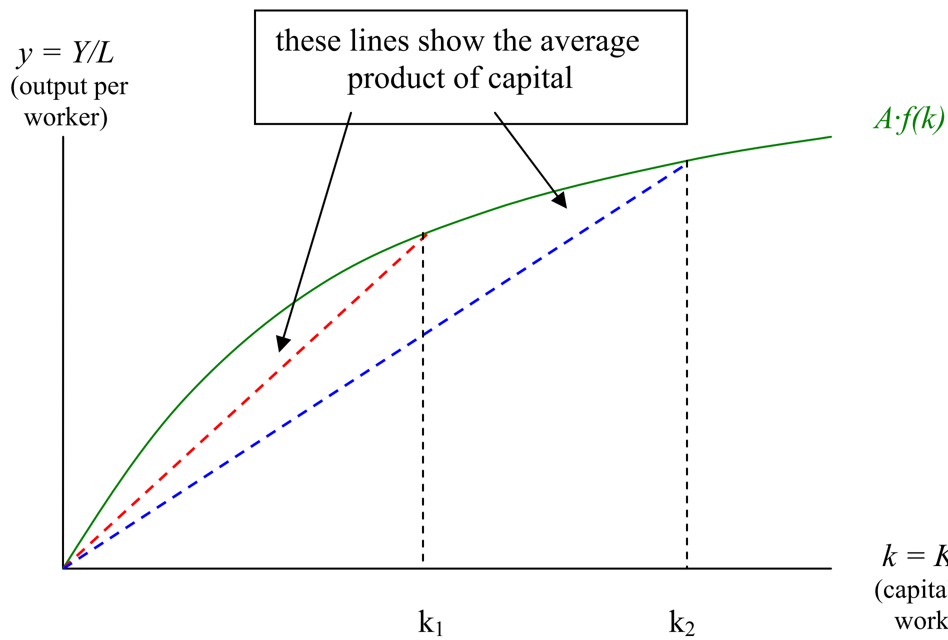


Figure 3.8

The Average Product of Capital

The graph shows the production function for output per worker, $y = Y/L$, versus capital per worker, $k = K/L$, as in Figure 3.7. The slope of a straight line from the origin to the production function gives the average product of capital, y/k, at the associated value of k. As k rises, for a given technology level, A, the average product of capital falls. For example, the slope of the dashed red line, which corresponds to $k_1 = K_1/L$, is greater than that of the dashed blue line, which corresponds to $k_2 = K_2/L$, where k_2 is greater than k_1. Therefore, the production function exhibits diminishing average product of capital.

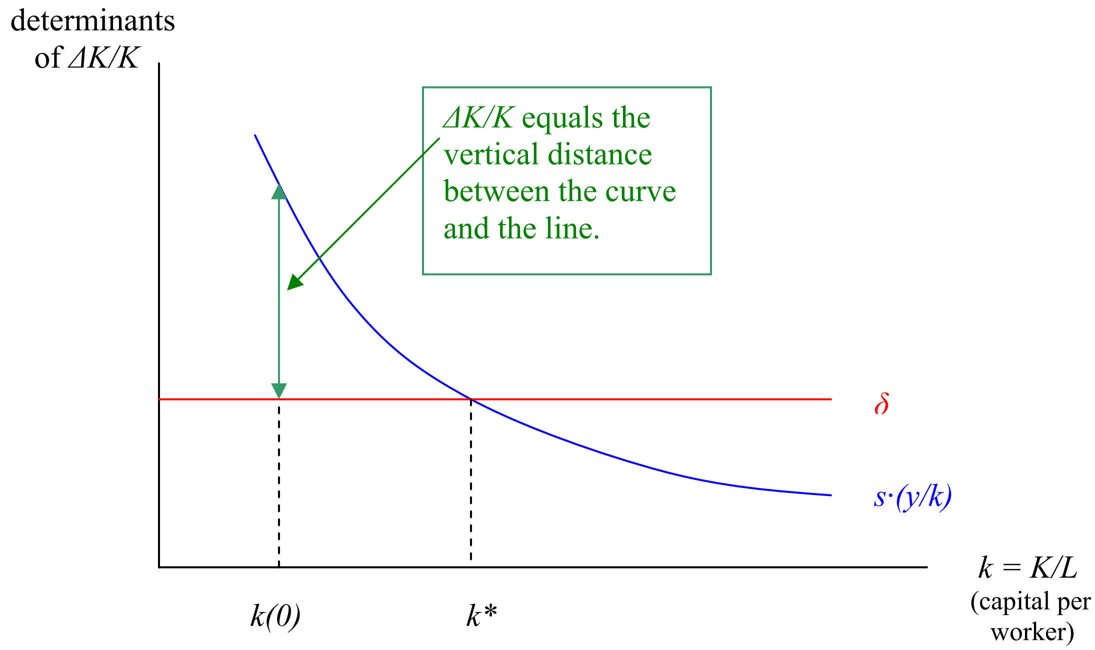


Figure 3.9

Determination of the Growth Rate of Capital in the Solow Model

The technology level, A, and labor input, L, are constant. The vertical axis plots the two determinants of the growth rate of capital, $\Delta K/K$, from the right-hand side of equation (3.9). $\Delta K/K$ equals the vertical distance between the negatively-sloped $s\cdot(y/k)$ curve (in blue) and the horizontal line at δ (in red). At the steady state, where $k = k^*$, the curve and line intersect, and $\Delta K/K = 0$. The initial capital per worker, $k(0)$, is assumed to be less than k^*. Therefore, when $k = k(0)$, $\Delta K/K$ is greater than zero and equal to the vertical distance shown by the green arrows.

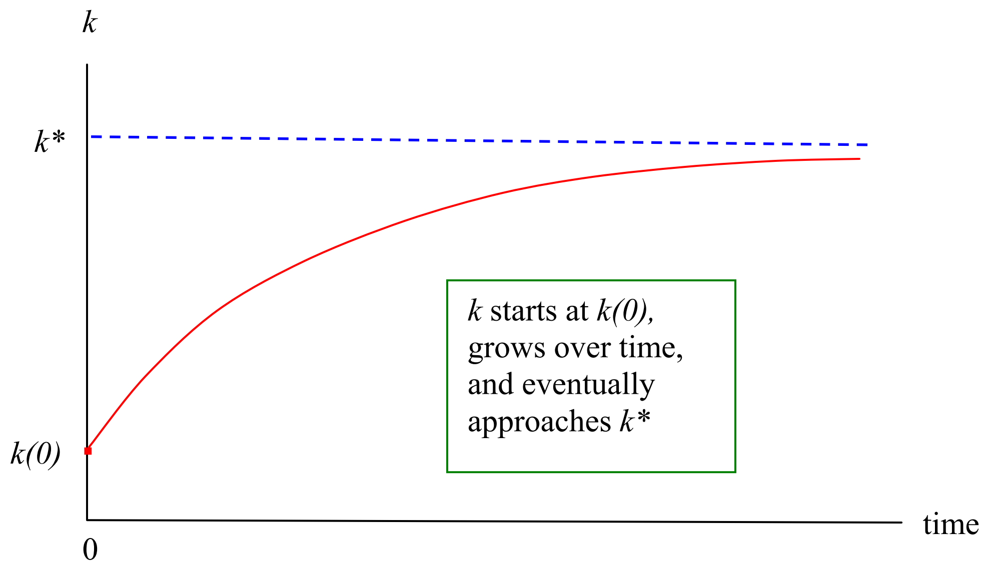


Figure 3.10

The Transition Path for Capital per Worker

In the Solow model, described by Figure 3.9, capital per worker, *k*, starts at *k(0)* and then rises over time. The growth rate of *k* slows down over time, and *k* gradually approaches its steady-state value, *k**. The transition path from *k(0)* to *k** is shown by the red curve. The dashed blue line shows the steady-state value, *k**.

January 12, 2005

Chapter 4

Working with the Solow Model

Now that we have constructed the Solow growth model we can put it into action by seeing how various economic changes affect growth in the short run and the long run. We begin by studying the effects from variations in the saving rate, s, the technology level, A, and the labor force, L. Then we explore the model's predictions about convergence, by which we mean a tendency for poor countries to catch up to rich ones.

We found in the Solow model that the equation for the growth rate of the capital stock is

(3.9) $$\Delta K/K = s\cdot(y/k) - \delta.$$

Thus far, we treated everything on the right-hand side as constant except for the average product of capital, y/k. We found that, in the transition to the steady state, the rise in capital per worker, k, led to a fall in y/k and, hence, to a fall in $\Delta K/K$. In the steady state, k was constant and, therefore, y/k was constant. In this situation, $\Delta K/K$ was constant and equal to zero.

We can substitute for real GDP per worker, y, from the production function,

(3.2) $$y = A\cdot f(k),$$

to get a revised version of equation (3.9):

(4.1) $$\Delta K/K = sA\cdot f(k)/k - \delta.$$

Up to now, we have assumed that the saving rate, s, and the technology level, A, were constant. Now we allow for changes in s and A. We also consider changes in labor input, L. We analyze the effects of these economic changes—variations in s, A, or L—on the two phases of the Solow model. What are the effects on the transition to the steady state, and what are the effects on the steady state? We can think of the first part as representing the short-run effects from an economic change and the second part as representing the long-run effects.

I. A Change in the Saving Rate

We now consider how differences in the saving rate, s, affect economic growth. As an example of differences in saving rates, we can compare nations in which the residents regularly save at a high rate—such as Singapore and South Korea or some other East Asian countries—with places in which the residents typically save at a low rate—such as most countries in sub-Saharan Africa or Latin America. Some of the differences in saving rates result from government policies and some from cultural differences. The important point is that saving rates differ across societies and over time.

Figure 4.1 extends the Solow model from Figure 3.9 to consider two saving rates, s_1 and s_2, where s_2 is greater than s_1. Each saving rate determines a different curve for $s\cdot(y/k)$—notice that the one with s_2 lies above that for s_1. Recall that the growth rate of capital, $\Delta K/K$, equals the vertical distance between the $s\cdot(y/k)$ curve and the horizontal line, δ. Therefore, $\Delta K/K$ is higher at any capital per worker, k, when the saving rate is s_2 rather than s_1. Specifically, at $k(0)$, $\Delta K/K$ is higher when the saving rate is s_2.[1]

[1] We have assumed that $k(0)$ is such that $\Delta K/K$ is greater than zero for both saving rates.

For either saving rate, the growth rate of capital, *ΔK/K*, declines as capital per worker, *k*, rises above *k(0)*. When the saving rate is s_1, *ΔK/K* reaches zero when *k* attains the steady-state value k_1^* in Figure 4.1. However, at k_1^*, *ΔK/K* would still be greater than zero if the saving rate were higher, in particular, if the saving rate equaled s_2. (See the figure.) Thus, if the saving rate is s_2, *k* rises beyond k_1^*—it increases until it reaches the steady-state value k_2^*, which is greater than k_1^*.

Using algebra

We can determine k^* algebraically from

(3.14) $$A{\cdot}f(k^*)/k^* = \delta/s.$$

An increase in *s* lowers δ/s on the right-hand side. Hence, the left-hand side must be lower, and this reduction can occur only through a decrease in the average product of capital, $A{\cdot}f(k^*)/k^*$. We know from Figure 3.8 that, if *A* is fixed, a decrease in the average product of capital requires an increase in capital per worker, *k*. Therefore, an increase in *s* raises k^*.

To summarize, in the short run, an increase in the saving rate, *s*, raises the growth rates of capital and real GDP. These growth rates remain higher during the transition to the steady state. In the long run, the growth rates of capital and real GDP are the same—zero—for any saving rate. In this long-run or steady-state situation, a higher saving rate leads to higher steady-state capital and real GDP per worker, k^* and y^*, not to changes in the growth rates (which remain at zero).[2]

Consumption in the Solow model

Since income equals consumption plus saving,

$$Y = C + sY,$$

an increase in saving means, for given income, that consumption must fall. However, since higher saving leads in the long run to higher income, consumption may increase in the long run. Here, we consider what the Solow model tells us about the effect of saving on consumption in the long run.

An increase in the saving rate, s, raises the steady-state capital and real GDP per worker, k^* and y^*. The rise in real GDP per worker leads to an increase in the typical person's real income. However, people care about their consumption, not about their income, *per se*. Thus, we want to know how a rise in the saving rate affects steady-state consumption per person.

Let c represent consumption per worker. Consumption per person is then

$$\text{consumption per person} = (\text{consumption per worker}) \times (\text{workers/population}).$$

[2] One important extension of the Solow model—carried out by David Cass and Tjalling Koopmans in the mid 1960s—was to allow households to choose the saving rate, *s*. To study this choice, we have to analyze how households determine consumption at different points in time. We defer this analysis to chapter 7.

Since workers/population equals the labor-force participation rate, which we have assumed to be constant, consumption per person always moves along with consumption per worker, c.

Since consumption equals the income that is not saved, and the saving per worker in the steady state is sy^*, we have that

(4.2) $$c^* = y^* - sy^*,$$

where c^* is the steady-state value of c. In the steady state, where the capital stock is constant, we know that saving, sY^*, equals gross investment, I^*, which just covers the depreciation of the capital stock, δK^*. Therefore, the steady-state saving per worker, sy^*, is

$$sy^* = \delta k^*.$$

If we substitute δk^* into equation (4.2) to replace sy^*, we get the result for c^*:

(4.3) $$c^* = y^* - \delta k^*.$$

We know that a rise in the saving rate, s, raises k^*, say, by the amount Δk^*. The change in c^* follows from equation (4.3) as

$$\Delta c^* = \Delta y^* - \delta \cdot \Delta k^*.$$

We can compute Δy^* by noting that it must equal Δk^* multiplied by the marginal product of capital, MPK:

$$\Delta y^* = MPK \cdot \Delta k^*.$$

Therefore, if we substitute $MPK \cdot \Delta k^*$ for Δy^*, we get that the change in c^* is given by

$$\Delta c^* = MPK \cdot \Delta k^* - \delta \cdot \Delta k^*$$

$$= (MPK - \delta) \cdot \Delta k^*.$$

Hence, Δc^* is greater than 0 as long as MPK is greater than δ.[3] Economists have found that the marginal product of capital, MPK, is well above the depreciation rate, so we can be confident that a rise in the saving rate increases steady-state consumption per worker, c^*. Therefore, a higher saving rate also raises steady-state consumption per person.

The positive effect of the saving rate, s, on steady-state consumption per person does not necessarily mean that the typical person is better off by saving more. In order to achieve the higher steady-state capital per worker, k^*, people have to save more during the transition to the steady state. Hence, the levels of consumption per person during the transition have to be reduced. Thus, there is a tradeoff—less consumption per person during the transition and more consumption per person in the long run. Whether the typical person is better or worse off depends on, first, how much consumption is gained in the long run in comparison with consumption lost during the transition and, second, on how patient people are about deferring consumption.

[3] This type of result was first presented by Edmund Phelps (1961).

II. A Change in the Technology Level

We have assumed, thus far, that the technology level, A, was constant. In reality, technology varies over time and across locations. For examples of improvements in technology over time, we can think of the introductions of electric power, automobiles, computers, and the Internet. For differences across locations, we can think of businesses in advanced countries, such as the United States and Western Europe, as having better access than their counterparts in poor countries to leading technologies. To assess the influences from differences in technologies, we begin by considering the effects in the Solow model from a change in the technology level, A.

The formula for the growth rate of the capital stock is

$$\Delta K/K = sA \cdot f(k)/k - \delta, \tag{4.1}$$

where $A \cdot f(k)/k$ is the average product of capital, y/k. Note that a higher A means that y/k is higher at a given value of k.

Figure 4.2 compares two levels of technology, A_1 and A_2, where A_2 is greater than A_1. Each technology level corresponds to a different curve for $s \cdot (y/k) = sA \cdot f(k)/k$. The curve with the higher technology level, $sA_2 \cdot f(k)/k$, lies above the other one, $sA_1 \cdot f(k)/k$. Notice that the positions of the two curves are similar to those in Figure 4.1, which considered two values of the saving rate, s. Hence, the formal analysis of a change in A is similar to that for a change in s.

At the initial capital per worker, $k(0)$, in Figure 4.2, the growth rate of the capital stock, $\Delta K/K$, is higher with the higher technology level, A_2, than with the lower one, A_1. In both cases, $\Delta K/K$ declines over time. For the lower technology level, $\Delta K/K$ falls to zero when capital per worker, k, attains the steady-state value k_1^*. For the higher technology level, k rises beyond k_1^* to reach the higher steady-state value k_2^*. Thus, an increase in A results in a higher $\Delta K/K$ over the transition period. In the long run, $\Delta K/K$ still falls to zero, but the steady-state capital per worker, k^*, is higher. That is, k_2^* is greater than k_1^*.

An increase in the technology level, A, raises the steady-state real GDP per worker, $y^* = A \cdot f(k^*)$, for two reasons. First, an increase in A raises real GDP per worker, y, for given capital per worker, k. Second, the steady-state capital per worker, k^*, is higher when A is higher. On both counts, an increase in A raises y^*.

To summarize, in the short run, an increase in the technology level, A, raises the growth rates of capital and real GDP. These growth rates remain higher during the transition to the steady state. In the long run, the growth rates of capital and real GDP are the same—zero—for any technology level. In this long-run or steady-state situation, a higher technology level leads to higher steady-state capital and real GDP per worker, k^* and y^*, not to changes in the growth rates (which remain at zero).

Using algebra

We can derive the effect of A on k^* algebraically from

(3.14) $$A \cdot f(k^*)/k^* = \delta/s.$$

An increase in A does not affect the right-hand side, δ/s. Therefore, the steady-state capital per worker, k^*, must adjust on the left-hand side to keep the steady-state average product of capital, $A \cdot f(k^*)/k^*$, the same. Since the increase in A raises this average product, k^* must change in a way that reduces the average product. As we know from Figure 3.8, a reduction in the average product of capital requires an increase in k^*. Therefore, an increase in A raises k^*.

III. A Change in the Labor Force

We have assumed, thus far, that the labor force and, hence, labor input, L, were constant. More realistically, the labor force varies across countries and over time. Here, we begin the study of variations in the labor force by considering the effects in the Solow model from different levels of L.

In the long run, the most important source of changes in labor input, L, is population growth. We postpone this important topic to the next section. In the short run, changes in L can result from shifts in the labor force. For example, the labor force could decline precipitously due to an epidemic of disease. An extreme case from the mid 1300s is the bubonic plague or Black Death, which is estimated to have killed about 20% of the European population. The potential loss of life due to the ongoing AIDS epidemic in Africa may turn out to be similar. In these examples, physical capital would not change, and the initial capital per worker, $k(0) = K(0)/L$, would rise sharply.

Another source of dramatic decrease in the labor force involves wartime casualties. However, since wartime tends also to destroy physical capital, the effect on capital per worker depends on the particular circumstances. Migration can also change the labor force dramatically in the short run. One example is the Mariel boatlift of over 100,000 Cuban refugees mostly to Miami in 1980. Another example is the dramatic in-migration to Portugal in the mid 1970s from its citizens who had been residing in African colonies. When these colonies became independent, many residents returned to Portugal, and this inflow added roughly 10% to the domestic Portuguese population.

Figure 4.3 illustrates the effects from a one-time increase in labor input from L_1 to L_2. We assume that the initial stock of capital, $K(0)$, is the same in both cases. Therefore, the increase in labor input from L_1 to L_2 leads to a decrease in the initial capital per worker, $k(0) = K(0)/L$. Specifically, $k(0)_2$ is less than $k(0)_1$ in the graph.

Note that the change in labor input, L, does not affect the curve for $s\cdot(y/k)$. However, the reduction in $k(0)$ from $k(0)_1$ to $k(0)_2$—due to the increase in L—leads to a higher value of $s\cdot(y/k)$ along the curve. That is, the reduction in capital per worker, k, raises the average product of capital, y/k. (See Figure 3.8.) The rise in y/k causes an increase in the growth rate of the capital stock, $\Delta K/K$. We can see this result in Figure 4.3 because the vertical distance between the $s\cdot(y/k)$ curve and the δ line is greater at $k(0)_2$ (the red arrows) than at $k(0)_1$ (the blue arrows). The growth rate $\Delta K/K$ then remains higher during the transition to the steady state. However, $\Delta K/K$ still declines toward its long-run value of zero. Moreover, the steady-state capital per worker, k^*, is the same when labor input is L_2 as it was when labor input was L_1. Since $k^* = K^*/L$ is the same, it must be that K^* rises in the same proportion as the increase in L. For example, if L_2 is twice as large as L_1, K^* is twice as large as it was initially.

To summarize, in the short run, an increase in labor input, L, raises the growth rates of capital and real GDP. These growth rates remain higher during the transition to the steady state. In the long run, the growth rates of capital and real GDP are the same—zero—for any quantity of labor input, L. Moreover, the steady-state capital and real GDP per worker, k^* and y^*, are the same for any L. In the long run, an economy with twice as much labor input has twice as much capital and twice as much real GDP.

Using algebra

We can again work out the steady-state results algebraically from the condition

(3.14) $$A \cdot f(k^*)/k^* = \delta/s.$$

Since A, s, and δ are all constant, the equation tells us that the steady-state capital per worker, k^*, does not change when L changes. Therefore, an increase in L does not affect k^*.

IV. Population Growth

We have already discussed one-time changes in the labor force. However, in the long run, the major source of increase in the labor force is **population growth**, that is, rises over time in the overall population. Population growth rates vary across countries and over time. In the United States, the population growth rate has been about 1% per year for several decades. In many Western European countries, population growth rates fell from around 1% per year in the 1960s to roughly zero in 2000. In China and India, population growth rates declined from over 2% per year in the 1960s to recent values between 1 and 1-1/2% per year. Many low-income countries still have population growth rates above 2% per year. However, there has been a worldwide tendency for population growth rates to decline over time.

A. The Solow model with population growth

Now we extend the Solow model to include population growth. Suppose that population is growing at the constant rate n, where n is a positive number ($n > 0$). Recall that the relation between the labor force and population is

$$labor\ force = (labor\ force/population) \times population,$$

where (labor force/population) is the labor-force participation rate. Since we assume that the labor-force participation rate is constant, we can see that the labor force grows at the same rate as population, that is, at rate n. Since we do not consider changes in hours worked per year, we still have that labor input, L, equals the labor force. Therefore, L also grows at the rate n.

Thus far, we have considered capital and real GDP per worker, k and y. We can translate these amounts per worker into amounts per person by using the equations:

$$capital\ per\ person = (capital\ per\ worker) \times (workers/population),$$

$$real\ GDP\ per\ person = (real\ GDP\ per\ worker) \times (workers/population).$$

Note that (workers/population) is the same as (labor force/population), which is the labor-force participation rate. If this participation rate is one-half, as in the recent U.S. data, capital per person is one-half of capital per worker, and real GDP per person is one-half of real GDP per worker.

Our previous analysis considered growth in the capital stock and real GDP at the rates $\Delta K/K$ and $\Delta Y/Y$, respectively. Since labor input, L, was constant, these growth rates also determined the growth rates of capital and real GDP per worker, $k = K/L$ and $y = Y/L$. For example, if L is constant, growth in K at the rate $\Delta K/K$ implies growth in $k = K/L$ at the same rate.

Suppose now that the labor force, L, and population grow at the rate n, so that

(4.4) $$\Delta L/L = n > 0.$$

If K were fixed, growth in L at the rate n would mean that capital per worker, $k = K/L$, would decline at the rate n. For example, with a fixed K, growth in the number of workers at 1% per year would mean that the quantity of capital available to each worker would fall by 1% per year. More generally, we have the formula

growth rate of capital per worker = growth rate of capital – growth rate of labor

$$\Delta k/k = \Delta K/K - \Delta L/L$$

(4.5) $$\Delta k/k = \Delta K/K - n.$$

Hence, for any growth rate of capital, $\Delta K/K$, a higher growth rate of labor, $\Delta L/L = n$, means that, over time, less capital is available for each worker. Using the same reasoning, the growth rate of real GDP per worker, $\Delta y/y$, falls short of the growth rate of real GDP, $\Delta Y/Y$, by the growth rate n of the number of workers:

growth rate of real GDP per worker = growth rate of real GDP – growth rate of labor

$$\Delta y/y = \Delta Y/Y - \Delta L/L$$

(4.6) $$\Delta y/y = \Delta Y/Y - n.$$

Recall that the growth-accounting equation is

(3.3) $$\Delta Y/Y = \Delta A/A + \alpha\cdot(\Delta K/K) + (1-\alpha)\cdot(\Delta L/L).$$

We continue to assume that the technology is constant, so that $\Delta A/A = 0$. However, we now have from equation (4.4) that $\Delta L/L = n$. If we substitute $\Delta A/A = 0$ and $\Delta L/L = n$ into equation (3.3), we get

$$\Delta Y/Y = \alpha\cdot(\Delta K/K) + (1\text{-}\alpha)\cdot n.$$

If we substitute this result for $\Delta Y/Y$ into equation (4.6), we get an equation for the growth rate of real GDP per worker:

$$\begin{aligned}\Delta y/y &= \Delta Y/Y - n \\ &= [\alpha\cdot \Delta K/K + (1\text{-}\alpha)\cdot n] - n \\ &= \alpha\cdot \Delta K/K + \cancel{n} - \alpha n - \cancel{n} \\ &= \alpha\cdot \Delta K/K - \alpha n \\ &= \alpha\cdot(\Delta K/K - n).\end{aligned}$$

Equation (4.5) tells us that the term on the right-hand side, $\Delta K/K - n$, equals the growth rate of capital per worker, $\Delta k/k$. Therefore, the result simplifies to

(4.7) $$\Delta y/y = \alpha\cdot(\Delta k/k).$$

Hence, the growth rate of real GDP per worker is the fraction α of the growth rate of capital per worker.

The presence of population growth does not change the formula for the growth rate of the capital stock:

(3.9) $$\Delta K/K = s\cdot(y/k) - \delta.$$

If we substitute $\Delta K/K = s\cdot(y/k) - \delta$ into equation (4.5), we can compute the growth rate of capital per worker, k:

$$\Delta k/k = s\cdot(y/k) - \delta - n,$$

or, if we combine δ and n, as

Key equation (Solow growth model with population growth):

(4.8) $$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

Equation (4.8) is a key result, which will allow us to assess the effects of population growth in the Solow model. To carry out this analysis, we need only to revise the horizontal line for depreciation, δ, in Figure 3.9. This line has to be modified to add the rate of population growth, n. Figure 4.4 shows the revised graph. Note that the new horizontal line is at $\delta+n$.

The new term, $\delta+n$, can be viewed from equation (4.8) as the **effective depreciation rate** for capital per worker, k. The rate δ refers to literal depreciation—it determines how the capital stock, K, would decline over time if saving were zero ($s = 0$) and labor input, L, were constant ($n = 0$). By analogy, the rate of population growth, n, determines how $k = K/L$ would fall over time due to rising L if saving were zero and δ were zero. Another way to say this is that n determines how much of investment is needed to provide the growing labor force, L, with capital to work with.

In Figure 4.4, the vertical distance between the $s\cdot(y/k)$ curve and the horizontal line at $\delta+n$ determines the growth rate of capital per worker, $\Delta k/k$, from equation (4.8). We assume that, at the initial capital per worker, $k(0)$, the distance is positive, so that $\Delta k/k$ is greater than zero. Then, as before, the increases in k lead over time to decreases in the average product of capital, y/k. These declines in y/k lead to decreases in $\Delta k/k$. (See equation [4.8].) Eventually, the economy reaches the steady-state capital per worker k^*. At this point, $\Delta k/k = 0$, and k stays constant.

In the steady state, capital per worker, k, equals the constant k^*. Therefore, equation (3.2) implies that real GDP per worker is constant at $y^* = A \cdot f(k^*)$. Since $\Delta k/k = 0$ and $\Delta y/y = 0$, equations (4.5) and (4.6) imply that the levels of capital and real GDP each grow in the steady state at the rate of population growth:

(4.9) $$(\Delta K/K)^* = (\Delta Y/Y)^* = n.$$

Therefore, if n is greater than zero, the growth rates of capital and real GDP are now greater than zero in the steady state.

Since $\Delta k/k = 0$ in the steady state, the right-hand side of equation (4.8) must equal zero:

$$s \cdot (y^*/k^*) - (\delta + n) = 0.$$

Therefore, if we move $\delta+n$ to the right-hand side, we get

(4.10) $$s \cdot (y^*/k^*) = \delta + n.$$

If we multiply through by k^*, we have

$$sy^* = \delta k^* + nk^*.$$

The left-hand side is saving per worker in the steady state. If $n = 0$, this saving just covers depreciation per worker, δk^*, which is the first term on the right-hand side. In this case, net investment is zero. However, if n is greater than zero, the steady-state saving per worker also includes the term nk^*. This term is the saving required in the steady state to provide the growing labor force with capital to work with. Because of this term, net investment is greater than zero in the steady state. This net investment allows the capital stock to grow at the rate $(\Delta K/K)^* = n$ (see equation [4.9]).

Using algebra

Algebraically, we can determine k* from the condition

$$(4.10) \qquad s\cdot(y^*/k^*) = \delta+n.$$

If we substitute $y^* = A\cdot f(k^*)$ from equation (3.2) and divide through by s, we get

$$(4.11) \qquad A\cdot f(k^*)/k^* = (\delta+n)/s.$$

This formula can be used to determine the steady-state capital per worker, k^*. This equation is the same as equation (3.14), except that the depreciation rate, δ, has been replaced by the effective depreciation rate, $\delta+n$.

B. A change in the population growth rate

We now consider how changes in the population growth rate, n, affect economic growth. Figure 4.5 compares two rates of population growth, n_1 and n_2, where n_2 is greater than n_1. Each population growth rate corresponds to a different line for the effective depreciation rate, $\delta+n$. Notice that the one for n_2 lies above that for n_1. The $s\cdot(y/k)$ curve is the same in both cases. Recall that the growth rate of capital per worker, $\Delta k/k$, equals the vertical distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line. Therefore, $\Delta k/k$ is lower at any capital per worker, k, when the population growth rate is n_2 rather than n_1. Specifically, at $k(0)$, $\Delta k/k$ is lower when the population growth rate is n_2. We can also see this result from equation (4.8). The reason that $\Delta k/k$ is lower when n is higher is that a larger portion of saving goes to providing the growing labor force, L, with capital to work with.

For either population growth rate, the growth rate of capital per worker, $\Delta k/k$, declines as capital per worker rises above $k(0)$. When the population growth rate is n_2, $\Delta k/k$ reaches zero when k attains the steady-state value k_2^* in Figure 4.5. However, at k_2^*, $\Delta k/k$ would still be greater than zero if the population growth rate were lower, in particular, if it equaled n_1. Thus, if the population growth rate is n_1, capital per worker rises beyond k_2^*—k increases until it reaches the steady-state value k_1^*, which is greater than k_2^*.

To summarize, in the short run, an increase in the population growth rate, n, lowers the growth rates of capital and real GDP per worker. These growth rates continue to remain lower during the transition to the steady state. In the long run, the growth rates of capital and real GDP per worker are the same—zero—for any n. In this long-run or steady-state situation, a higher n leads to lower steady-state capital and real GDP per worker, k^* and y^*, not to changes in the growth rates, $\Delta k/k$ and $\Delta y/y$ (which remain at zero). However, a change in n does affect the steady-state growth rates of the levels of capital and real GDP, $\Delta K/K$ and $\Delta Y/Y$. An increase in n by 1% per year raises $\Delta K/K$ and $\Delta Y/Y$ by 1% per year.

Using algebra

We can find the effect of n on k^* algebraically from the condition

(4.11) $$A\cdot f(k^*)/k^* = (\delta+n)/s.$$

An increase in n raises the right-hand side of the equation. Hence, the steady-state average product of capital, $A\cdot f(k^*)/k^* = y^*/k^*$, has to rise on the left-hand side.

Because of the diminishing average product of capital (Figure 3.8), this change requires a decrease in k^*. Therefore, an increase in n reduces k^*.

We can also see from Figure 4.5 and equation (4.8) that an increase in the depreciation rate, δ, would affect the steady-state capital per worker in the same way as an increase in the population growth rate, n. This result follows because the model depends on the effective rate of depreciation, $\delta+n$, which can rise either from an increase in n or an increase in δ. The kind of analysis that we carried out for an increase in n tells us that an increase in δ would lower the growth rates of capital and real GDP per worker in the short run. In the long run or steady state, an increase in δ leads to lower capital and real GDP per worker, k^* and y^*, not to changes in the growth rates—which remain at zero.[4]

Endogenous population growth

Our analysis treated the population growth rate, n, as exogenous—determined outside of the model. However, at least since the writings of Thomas Malthus, economists have argued that population growth responds to economic variables. Malthus was a British economist and minister who wrote his *Essay on Population* in 1798. He argued that an increase in real income per person raised population growth by improving life expectancy, mainly through better nutrition

[4] One difference is that an increase in n raises the steady-state growth rates of the levels of capital and real GDP, $(\Delta K/K)^*$ and $(\Delta Y/Y)^*$, whereas an increase in δ does not affect these steady-state growth rates.

but also through improved sanitation and medical care. Another influence, Malthus thought, was that higher income encouraged greater fertility. He thought that birth rates would rise as long as real income per person exceeded the **subsistence level**, which is the amount needed to provide the basic necessities of life.

We can incorporate Malthus's ideas about population growth into the Solow model. In Figure 4.4, for a given rate of population growth, n, the economy approaches a steady-state capital per worker, k^*, and a corresponding real GDP per worker, $y^*=A\cdot f(k^*)$. For a given labor-force participation rate, the real GDP per worker determines the real GDP per person.

If real GDP per person and, hence, household real income per person were above the subsistence level, Malthus believed that the population growth rate would rise. Figure 4.5 showed the effect from a rise in the population growth rate. This change lowered the steady-state capital and real GDP per worker. According to Malthus, this process would continue until the steady-state real income per person fell to the subsistence level.

Malthus's view on the relation between real income per person and life expectancy is reasonable. Data across countries show that higher real GDP per person matches up closely with higher life expectancy at birth.[5] However, Malthus's idea about fertility seems unreasonable. At least in the cross-country data since 1960, higher real GDP per person matches up with lower fertility.[6] In fact, this relation is so strong that a higher real GDP per person also matches up with a lower population growth rate, even though countries with higher real GDP per person have higher life expectancy.

We can modify the Solow model to include Malthus's idea that population growth is endogenous. However, contrary to Malthus, we should assume a negative effect of real GDP per person—and, hence, of capital per worker, k—on the population growth rate, n.

We allowed for population growth, n, in equation (4.8):

(4.8) $$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

During the transition to the steady state, a rise in k reduced the average product of capital, y/k, and thereby decreased the growth rate, $\Delta k/k$. Now we have that a rise in k also lowers n. This change raises $\Delta k/k$ and, thereby, offsets the effect from a reduced average product of capital. Hence, a declining population

[5] Although this relation is suggestive, it does not prove that the causation is from higher real income per person to greater life expectancy, rather than the reverse. In fact, both directions of causation seem to be important.

[6] This relation does not prove that the causation is from higher real income per person to lower fertility, rather than the reverse. In fact, the reverse effect is predicted by the Solow model. If a society chooses, perhaps for cultural reasons, to have higher fertility and, hence, higher population growth, the model predicts lower steady-state real GDP per worker. In practice, both directions of causation seem to be important.

growth rate is one reason why rich societies can sustain growing capital and real GDP per worker for a long time.

V. Convergence

One of the most important questions about economic growth is whether poor countries tend to converge or catch up to rich countries. Is there a systematic tendency for low-income countries like those in Africa to catch up to the rich OECD countries? We start our answer to this question by seeing what the Solow model says about **convergence**. Then we look at the facts on convergence and see how these match up with the Solow model.

A. Convergence in the Solow model

To study convergence, we focus on the transition for capital per worker, k, as it rises from its initial value, $k(0)$, to its steady-state value, k^*. In Figure 3.10, we can see that k^* works like a target or magnet for k during the transition. Therefore, any economic change that affects k^* will alter the transition path for k. Thus, we now summarize in the form of a function what we have learned about the determinants of k^*:

$$k^* = k^*(\underset{(+)}{s}, \underset{(+)}{A}, \underset{(-)}{n}, \underset{(-)}{\delta}, \underset{(0)}{L}). \tag{4.12}$$

The sign below each variable indicates the effect on k^*. Recall that k^* rises with the saving rate, s, and the technology level, A, and falls with the population growth rate, n, and the depreciation rate, δ. Finally, k^* is independent of the labor force, L. Table 4.1 summarizes these findings.

Table 4.1	
Effects on Steady-State Capital per Worker, k^*	
Variable	**Effect on k^***
saving rate, s	increase
technology level, A	increase
depreciation rate, δ	decrease
population growth rate, n	decrease
labor force, L	no effect

Note: These results come from equation (4.12).

To apply the Solow model to convergence, we have to allow for more than one economy. In making this extension, we continue to assume that the economies are independent of each other. Specifically, they do not engage in international trade in goods and services or in financial assets.

Think now of two economies, 1 and 2, and suppose that they start with capital per worker of $k(0)_1$ and $k(0)_2$, respectively, where $k(0)_1$ is less than $k(0)_2$. Thus, economy 2 is initially more advanced in the sense of having higher capital and real GDP per worker. Imagine that each economy has the same values for the determinants of k^* listed in Table 4.1, so that they have the same steady-state capital per worker, k^*.

We show this situation in Figure 4.6, which has been adapted from Figure 4.4. The only difference between the two economies is that one starts at $k(0)_1$ and the other at $k(0)_2$. Therefore, the differences in the transition paths of k depend only on the differences in these starting values. The graph in Figure 4.6 shows that the vertical distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line is greater at $k(0)_1$ than at $k(0)_2$. That is, the distance marked with red arrows is greater than that marked with blue arrows. Therefore, the growth rate $\Delta k/k$ is higher initially for economy 1 than economy 2. Because k grows at a faster rate in economy 1, economy 1's level of k converges over time toward economy 2's.

Figure 4.7 shows the transition paths of capital per worker, k, for economies 1 and 2. Note that $k(0)_1$ is less than $k(0)_2$, but k_1 gradually approaches k_2. (At the same time, k_1 and k_2 both gradually approach k^*.) Thus, economy 1 converges toward economy 2 in terms of the levels of k.

We can also express the results in terms of real GDP per worker, y. The capital per worker, k, determines y from the production function as $y = A\cdot f(k)$ (equation [3.3]). Since economy 1 starts out with lower capital per worker, $k(0)$, it must start out also with lower real GDP per worker—$y(0)_1$ is less than $y(0)_2$. The growth rate of real GDP per worker relates to the growth rate of capital per worker from

$$(4.7) \qquad \Delta y/y = \alpha\cdot(\Delta k/k).$$

We showed in Figure 4.6 that $\Delta k/k$ was higher initially in economy 1 than economy 2. Therefore, $\Delta y/y$ is also higher initially in economy 1. Hence, economy 1's real GDP per worker, y, converges over time toward economy 2's. The transition paths for y in the two economies look like the ones shown for k in Figure 4.7.

To summarize, the Solow model says that a poor economy—with low capital and real GDP per worker—grows faster than a rich one. The reason is the diminishing average product of capital, *y/k*. A poor economy, such as economy 1 in Figure 4.6, has the advantage of having a high average product of capital, *y/k*. This high average product explains why the growth rates of capital and real GDP per worker are higher in the economy that starts out behind. Thus, the Solow model predicts that poorer economies tend to converge over time toward richer ones in terms of the levels of capital and real GDP per worker.

B. Facts about convergence

The main problem with these predictions about convergence is that they conflict with the data for a broad group of countries. We already looked in Figure 3.3 at growth rates of real GDP per person from 1960 to 2000. To apply the Solow model to these data, we have to remember how to translate from amounts per worker to amounts per person. The formula for real GDP per person is again

real GDP per person = (real GDP per worker) x (workers/population).

The ratio workers/population is the labor-force participation rate, which we have assumed to be constant. For example, if the ratio is one-half, real GDP per person is one-half of real GDP per worker.

With this translation, we find that the Solow model predicts convergence for real GDP per person. Specifically, the model predicts that a lower level of real GDP per person would match up with a higher subsequent growth rate of real GDP per person.

Figure 4.8 uses the data for countries from Figure 3.3 to plot growth rates of real GDP per person from 1960 to 2000 against levels of real GDP per person in 1960. If the convergence prediction from the Solow model were correct, we would expect to find low levels of real GDP per person matched with high growth rates and high levels of real GDP per person matched with low growth rates. Instead, it is difficult to observe any pattern in the data—if anything, there is a slight tendency for the growth rate to rise with the level of real GDP per person.

The sample of countries included in Figure 4.8 is large; it includes the richest and poorest economies in the world. The convergence prediction of the Solow model fits better with the data if we limit the observations to economies that have similar economic and social characteristics. Figure 4.9 is the same as Figure 4.8, except that the sample is limited to 18 advanced countries.[7] For this limited sample, lower levels of real GDP per person in 1960 do match up, on average, with higher growth rates from 1960 to 2000. This pattern reflects especially the catching up of some of the initially poorer countries—Greece, Ireland, Portugal, and Spain—to the richer ones.[8]

Figure 4.10 shows an even clearer pattern of convergence for a still more homogenous group of economies—the states of the United States. The figure plots the average growth rate of personal income per person from 1880 to 2000 against the level of

[7] These 18 were picked from the 20 founding members in 1960 of the rich countries' club, the OECD (Organization for Economic Cooperation and Development). Germany was not included because the unification of East and West in 1990 makes it difficult to consider the full period 1960-2000. Turkey was excluded because its OECD membership was dictated by political considerations, rather than its economic and social characteristics.

[8] Figure 4.9 excludes Japan because it was not a founding member of the OECD. If Japan were included, it would fit the pattern shown in the graph—a relatively low starting real GDP per person matching up with a relatively high growth rate.

personal income per person in 1880.[9] This graph shows a dramatic tendency for the initially poorer states to grow faster than the initially richer states over the 120 years after 1880. Moreover, this convergence tendency does not reflect only the recovery of the southern states, which were defeated during the U.S. Civil War (1861-1865). The convergence pattern applies almost as strongly if we examine economic performance within any of the four main regions—northeast, south, midwest, and west. Results similar to those in Figure 4.10 have also been found for regions of other advanced countries.

From Figures 4.8-4.10, we learn that similar economies tend to converge, whereas dissimilar economies display no relationship between the level of real GDP per person and the growth rate. Thus, the convergence pattern is strongest for regions of advanced countries (Figure 4.10), next strongest among the group of rich countries (Figure 4.9), and weakest—in fact, absent—for the full worldwide sample of countries (Figure 4.8).

C. Conditional convergence in the Solow model

The Solow model's prediction of convergence seems to explain growth patterns in similar economies, and it seems to fail when we examine a dissimilar array of economies. Do these findings mean that the model is flawed? Are there changes we might make to improve its predictions?

To find the flaw and try to correct it, let's reexamine the Solow model. One key assumption was that the determinants of the steady-state capital per worker, k^*, were the same for all economies. This assumption is reasonable for similar economies but is less

[9] We use personal income because data on gross domestic product are unavailable for the individual U.S. states over the period since 1880.

reasonable when applied to a broad sample of countries having sharply different economic, political, and social characteristics. In particular, the assumption is unreasonable for the worldwide group of countries considered in Figure 4.8. To explain the lack of convergence in this setting, we have to allow for differences in the steady-state positions, k^*.

Suppose that countries differ with respect to some of the determinants of k^* in equation (4.12) and Table 4.1. For example, k^* could vary because of differences in saving rates, s, levels of technology, A, and population growth rates, n.[10] Figure 4.11 modifies Figure 4.6 to show how differences in saving rates affect convergence. Economy 1 has the saving rate s_1, and economy 2 has the higher saving rate s_2. We assume, as in Figure 4.6, that economy 1 has the lower initial capital per worker, that is, $k(0)_1$ is less than $k(0)_2$. Remember that the growth rate $\Delta k/k$ equals the vertical distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line. We see from the graph in Figure 4.11 that it is uncertain whether the distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line is greater initially for economy 1 or economy 2. The lower capital per worker, $k(0)$, for economy 1 tends to make the distance greater for economy 1, but the lower saving rate for economy 1 tends to make the distance smaller for economy 1. In the graph, these two forces roughly balance, so that $\Delta k/k$ is about the same for the two economies. That is, the distance marked with the blue arrows is similar to the one marked with the red arrows. Therefore, the poorer economy, economy 1, does not necessarily converge toward the richer economy, economy 2.

[10] Levels of population and, therefore, sizes of the labor force, L, vary greatly across countries, but L does not affect k^* in the model. The depreciation rate, δ, probably does not vary systematically across countries.

To get the result shown in Figure 4.11, we had to assume that the economy with the lower $k(0)$—economy 1—had the lower saving rate, that is, s_1 was less than s_2. This assumption makes sense because an economy with a lower saving rate has a lower steady-state capital per worker, k^*. In the long run, an economy's capital per worker, k, would be close to its steady-state value, k^*. Therefore, it is likely when we look at an arbitrary date, such as date 0, that $k(0)$ will be lower in the economy with the lower saving rate, that is, $k(0)$ would tend to be lower in economy 1 than in economy 2. Thus, the pattern that we assumed—a low saving rate matched with a low $k(0)$—is likely to apply in practice.

We get a similar result if we consider other reasons for differences in the steady-state capital per worker, k^*. Suppose, for example, that the two economies have the same saving rates but that economy 1 has a lower technology level than economy 2, that is, A_1 is less than A_2. In this case, the two curves for $s \cdot (y/k)$ again look as shown in Figure 4.11. (Why is this?) Therefore, it is again uncertain whether the vertical distance between the $s \cdot (y/k)$ curve and the $\delta+n$ line is greater for economy 1 or economy 2. The lower capital per worker, $k(0)$, for economy 1 tends to make the distance greater for economy 1, but the lower technology level for economy 1 tends to make the distance smaller for economy 1. As before, it is possible that the two forces roughly balance, so that $\Delta k/k$ is about the same for the two economies. Thus, once again, the poorer economy, economy 1, need not converge toward the richer economy, economy 2.

To get this result, we had to assume that the economy with the lower starting capital per worker, $k(0)$—economy 1—had the lower technology level. This assumption makes sense because an economy with a lower technology level has a lower steady-state capital per worker, k^*. Therefore, it is likely when we look at the two economies at an arbitrary date, such as date 0, that $k(0)$ will be lower in economy 1 than economy 2.

The same conclusions apply if we consider differences in population growth rates. In Figure 4.12, the two economies have the same saving rates and technology levels, but economy 1 has a higher population growth rate, that is, n_1 is greater than n_2. Hence, the $\delta+n$ line is higher for economy 1. It is again uncertain whether the vertical distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line is greater for economy 1 or economy 2. The lower capital per worker, $k(0)$, for economy 1 tends to make the distance greater for economy 1, but the higher population growth rate for economy 1 tends to make the distance smaller for economy 1. As in our other cases, it is possible that the two forces roughly balance, so that $\Delta k/k$ would be about the same in the two economies—the distance marked with the blue arrows is similar to the one marked with the red arrows. Thus, economy 1 again need not converge toward economy 2.

To get this result, we had to assume that the economy with the lower starting capital per worker, $k(0)$—economy 1—had the higher population growth rate. This assumption makes sense because an economy with a higher population growth rate has a lower steady-state capital per worker, k^*. Therefore, it is again likely when we look at the two economies at date 0 that $k(0)$ will be lower in economy 1 than economy 2.

Now let's try to generalize the conclusions from our three cases. In each case, economy 1 had a characteristic—lower saving rate, lower technology level, higher population growth rate—that led to a lower steady-state capital per worker, k^*. For a given starting capital per worker, $k(0)$, each of the three characteristics tended to make country 1's initial growth rate less than economy 2's. We see these effects in Figures

4.11 and 4.12. At a given $k(0)$, the vertical distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line is smaller if s or A is lower or if n is higher.

However, since economy 1 has lower k^*, it is also likely to have lower initial capital per worker, $k(0)$. The lower $k(0)$ tends to make economy 1 grow faster than economy 2—this is the convergence force shown in Figure 4.6. Thus, whether economy 1 grows faster or slower than economy 2 depends on the offset of two forces. The lower $k(0)$ generates faster growth in economy 1, but the lower k^* generates slower growth in economy 1. It is possible that the two forces roughly balance, so that the two economies grow at about the same rate. That is, we need not find convergence.

Figure 4.13 shows the transition paths of capital per worker, k, for the two economies. We assume that economy 1 starts with a lower capital per worker—$k(0)_1$ is less than $k(0)_2$—and also has a lower steady-state capital per worker—k_1^* is less than k_2^*. The graph shows that the capital per worker in each economy converges toward its own steady-state level—k_1 toward k_1^* and k_2 toward k_2^*. However, since k_1^* is less than k_2^*, k_1 does not converge toward k_2.

We can summarize our findings for growth rates in the form of an equation:

Key equation:

growth rate of capital per worker = function of initial and steady-state capital per worker

$$\Delta k/k = \varphi[\underset{(-)}{k(0)},\ \underset{(+)}{k^*}], \tag{4.13}$$

where φ represents the function that determines $\Delta k/k$. The minus sign under $k(0)$ means that, for given k^*, a decrease in $k(0)$ raises $\Delta k/k$.. The plus sign under k^* means that, for given $k(0)$, a rise in k^* increases $\Delta k/k$. We can interpret these effects from the perspective of our earlier equation for $\Delta k/k$:

(4.8) $$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

Since $y = A\cdot f(k)$ (equation [3.2]), we can rewrite this equation as

(4.14) $$\Delta k/k = sA\cdot f(k)/k - (\delta+n).$$

The negative effect of $k(0)$ in equation (4.13) corresponds in equation (4.14) to a higher initial average product of capital, $A\cdot f(k)/k$. The positive effect of k^* in equation (4.13) corresponds in equation (4.14) to a higher saving rate, s, a higher technology level, A, or a lower population growth rate, n.

One important result in equation (4.13) is that the negative effect of $k(0)$ on the growth rate, $\Delta k/k$, holds only in a conditional sense, for a given k^*. This pattern is called **conditional convergence**: a lower $k(0)$ predicts a higher $\Delta k/k$, conditional on k^*. In contrast, the prediction that a lower $k(0)$ raises $\Delta k/k$ without any conditioning is called **absolute convergence**.

Recall from Figure 4.8 that we do not observe absolute convergence for a broad group of countries. We can now see from equation (4.13) that we can use the Solow model to explain the lack of convergence in this diverse group. Suppose that some countries have low saving rates, low technology levels, or high population growth rates and, therefore, have low steady-state capital per worker, k^*. In the long run, capital per worker, k, will be close to k^*. Therefore, when we look at date 0, we tend to find that low values of $k(0)$ are matched with low values of k^*. A low value of $k(0)$ makes the growth rate of capital per worker, $\Delta k/k$, high, but a low value of k^* makes $\Delta k/k$ low. Thus, it is possible that the data will show little relation between the starting capital per worker, $k(0)$, and $\Delta k/k$. This pattern is consistent with the one found in Figure 4.8 for growth rates and levels of real GDP per person.

VI. Where Do We Stand with the Solow Model?

When we first considered convergence, we observed that the lack of absolute convergence for a broad group of countries, as in Figure 4.8, was a failing of the Solow model. Then we found that an extension of the model to consider conditional convergence explained this apparent failure. We show in the next chapter that conditional convergence allows us to understand many patterns of economic growth in the world.

Although the Solow model has many strengths, we should be clear about what the model does not explain. Most important is the failure to explain how real GDP per person grows in the long run—for example, at a rate of about 2% per year for well over a century in the United States and other advanced countries. In the model, capital per worker—and, hence, real GDP per worker and per person—are constant in the long run. Thus, a key objective of the next chapter is to extend the model to explain long-run economic growth.

Questions and Problems

Mainly for review

4.1. If labor input, L, doubles, why do we get the result that the steady-state capital stock, K^*, doubles? That is, Figure 4.3 implies that steady-state capital per worker, k^*, does not change. How does the result depend on constant returns to scale in the production function?

4.2. Does population growth, $n > 0$, lead to growth of output in the long run? Does it lead to growth of output per worker in the long run?

4.3. What is the meaning of the term convergence? How does absolute convergence differ from conditional convergence?

4.4. For 111 countries, Figure 4.8 shows that the growth rate of real per capita GDP from 1960 to 2000 bears little relation to the level of real GDP in 1960. Does this finding conflict with the Solow model of economic growth? How does this question relate to the concept of conditional convergence?

Problems for discussion

4.x. Variations in the saving rate

Suppose that the gross saving rate, s, can vary as an economy develops.

a. The equation for the growth rate of capital per worker, k, is given from equation (4.9) by

$$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

Is this equation still valid when s is not constant?

b. Suppose that s rises as an economy develops—that is, rich countries save at a higher rate than poor countries. How does this behavior affect the results about convergence?

c. Suppose, instead, that s falls as an economy develops—that is, rich countries save at a lower rate than poor countries. How does this behavior affect the results about convergence?

d. Which case seems more plausible—**b.** or **c.**? Explain.

4.x. Variations in the population growth rate

Suppose that the population growth rate, n, can vary as an economy develops.

a. The equation for the growth rate of capital per worker, k, is given from equation (4.9) by

$$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

Is this equation still valid when n is not constant?

b. Suppose that n falls as an economy develops—that is, rich countries have lower population growth than poor countries. How does this behavior affect the results about convergence?

c. Suppose, instead, that n rises as an economy develops—that is, rich countries have higher population growth than poor countries. How does this behavior affect the results about convergence?

d. Which case seems more plausible—**b.** or **c.**? Explain, giving particular attention to the views of Malthus about endogenous population growth.

Appendix

The rate of convergence

We assess here how fast convergence takes place in the Solow model. Figure 4.14 starts by reproducing the results from Figure 4.4. The effective depreciation line is at $\delta+n$, and the capital per worker starts at $k(0)$. The previous saving curve is now shown in blue as $s\cdot(y/k)^I$. The growth rate $\Delta k/k$ equals the vertical distance between the $s\cdot(y/k)^I$ curve and the $\delta+n$ line.

As stressed before, the source of convergence in the Solow model is the diminishing average product of capital, y/k. Recall that this average product is given by

$$y/k = A\cdot f(k)/k.$$

The tendency for the average product to fall as k increases is the reason that the $s\cdot(y/k)^I$ curve slopes downward. The slope of the curve determines how fast convergence occurs, and this slope will depend on the form of the function $f(k)/k$.

To understand the role of the slope of the $s\cdot(y/k)$ curve, Figure 4.14 includes a second saving curve, $s\cdot(y/k)^{II}$, shown in red. In comparison with the first curve, the second one has the same saving rate, s, and technology level, A, but a different form of the function $f(k)/k$. This different function means that the relation between k and y/k is different for curve II than for curve I. Specifically, at any value of k, the second curve does not slope downward as much as the first one. That is, the average product of capital, y/k, diminishes less rapidly with k in the second case than in the first one.

To ease the comparison, we set up the graph so that the two saving curves intersect the $\delta+n$ line at the same point. Hence, the steady-state capital per worker, k^*, is the same in the two cases. However, at the initial capital per worker, $k(0)$, the vertical distance between the saving curve and the $\delta+n$ line is larger in the first case than the second one. In the graph, the first distance equals the sum of the green and brown arrows, whereas the second distance equals the green arrows only. Therefore, at $k(0)$, the growth rate $\Delta k/k$ is higher in the first case. The higher growth rate means that k converges more rapidly toward the steady-state level, k^*, in the first case. Hence, we have shown that the rate of convergence is higher when the average product of capital diminishes more rapidly with k.

For a given technology level, A, the relation between the average product of capital, y/k, and k depends on the function $f(k)/k$. As an example, we can use the Cobb-Douglas production function, which we introduced in the appendix to the previous chapter. In this case, the function $f(k)$, given in equation (3.18), is k^{α}. Therefore, the average product of capital is

$$\begin{aligned} y/k &= A\cdot f(k)/k \\ &= Ak^{\alpha}/k \\ &= Ak^{\alpha}\cdot k^{-1} \\ &= Ak^{(\alpha-1)} \end{aligned}$$

$$y/k = Ak^{-(1-\alpha)}. \tag{4.15}$$

Since $0<\alpha<1$, the average product of capital, y/k, falls as k rises. The value of α determines how fast y/k falls as k rises. If α is close to 1, equation (4.15) says that y/k is nearly independent of k, and y/k falls only slowly as k rises. Curve II in Figure 4.14 is like this. Conversely, if α is close to zero, y/k falls quickly as k rises. Curve I is like this. Generally, the lower α the more quickly y/k falls as k rises.

To get a quantitative idea about the rate of convergence, consider the intermediate case in which $\alpha = 0.5$. In this case, the average product of capital is

$$y/k = Ak^{-(1/2)}$$

$$= A/\sqrt{k}\,,$$

that is, the average product of capital declines with the square root of k.

Recall that the growth rate of k is given by

(4.8) $$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

If we substitute $y/k = A/\sqrt{k}$, we get

(4.16) $$\Delta k/k = sA/\sqrt{k} - (\delta+n).$$

If we specify values for the saving rate, s, the technology level, A, the depreciation rate, δ, the rate of population growth, n, and the initial capital per worker, $k(0)$, we can use equation (4.16) to calculate the time path of k.[11]

Table 4.2 shows the answers. The calculations assume that the initial capital per worker, $k(0)$, is one-half of its steady-state value, k^*. (The value of $k[0]$ turns out not to affect the results.) The table reports the values of k/k^* and y/y^* that prevail after 5 years, 10 years, and so on. Note that it takes about 25 years—roughly a generation—to eliminate half of the initial gap between k and k^*. By analogy to radioactive decay in physics, we can define the time for half of the convergence to the steady state to occur as the **half-life**. Since the ratio k/k^* starts at 0.5 and the half-life of the convergence process is 25 years, the ratio reaches 0.75 in 25 years and 0.875 in 50 years. Hence, although capital per worker, k, converges toward k^*, the Solow model predicts that this process

[11] If we know s, A, δ, n, and $k(0)$, equation (4.16) determines k at the next point in time, $k(1)$. Then, once we know $k(1)$, we can use the equation again to calculate $k(2)$. Proceeding in this way, we can calculate $k(t)$ at any time t.

takes a long time. The same numerical results on half-lives turn out to apply to the adjustment of real GDP per worker, y, to its steady-state level, y^*.

If α is greater than 0.5, the average product of capital, y/k, declines more slowly as k rises, and the convergence to the steady state is less rapid. Therefore, the half-life is more than 25 years. Conversely, if α is less than 0.5, y/k declines more quickly as k rises, and the convergence to the steady state is more rapid. In this case, the half-life is less than 25 years.

Table 4.2
The Transition Path in the Solow Model

Year	k/k^*	y/y^*
0	0.50	0.71
5	0.56	0.75
10	0.61	0.78
15	0.66	0.81
20	0.71	0.84
25	0.74	0.86
30	0.78	0.88
35	0.81	0.90
40	0.83	0.91
45	0.86	0.93
50	0.88	0.94

Note: The table shows the solution of the Solow model for capital per worker, k, and real GDP per worker, y. The results are expressed as ratios to the steady-state values, k/k^* and y/y^*. The transitional behavior of k and y comes from equation (4.16), which assumes $y = A \cdot \sqrt{k}$. The calculations assume that k/k^* starts at 0.5 and uses the values $n = 0.01$ per year and $\delta = 0.05$ per year. The values of s, A, and L turn out not to affect the results.

Many interesting applications have been made for speeds of convergence and half-lives calculated from the Solow model. One implication involves the aftermath of the U.S. Civil War, which ended in 1865. The model says that the southern states from the defeated Confederacy would converge only slowly in terms of income per person to the richer northern states. (The war reduced the income per person in the south from roughly 80% of the northern level to 40%.) The quantitative prediction, which turns out to be accurate, is that the process would take more than two generations to be nearly complete.

Similarly, with the unification of Germany in 1990, the model predicts that the poor eastern parts from the formerly Communist East Germany would converge, but only slowly, to the richer western regions. (In 1990, the GDP per person of the eastern regions was about 1/3 of the western level.) This prediction for gradual convergence fits the data for the 1990s.

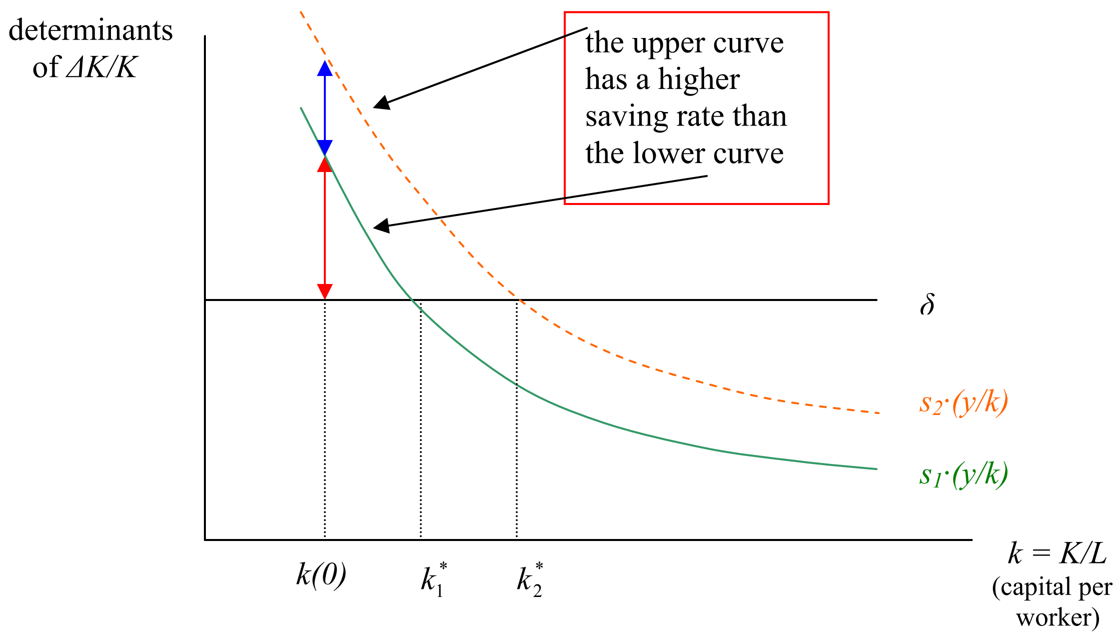


Figure 4.1

Effect of an Increase in the Saving Rate in the Solow Model

This graph comes from Figure 3.9. The curves for $s \cdot (y/k)$ are for the saving rates s_1 and s_2, where s_2 is greater than s_1. The growth rate of capital, $\Delta K/K$, is higher at any k when the saving rate is higher. For example, at $k(0)$, when the saving rate is s_1, $\Delta K/K$ equals the vertical distance shown by the lower red arrows. When the saving rate is s_2, $\Delta K/K$ equals the vertical distance that adds in the upper blue arrows. In the steady state, $\Delta K/K$ is zero, regardless of the saving rate. The higher saving rate yields a higher steady-state capital per worker, that is, k_2^* is greater than k_1^*.

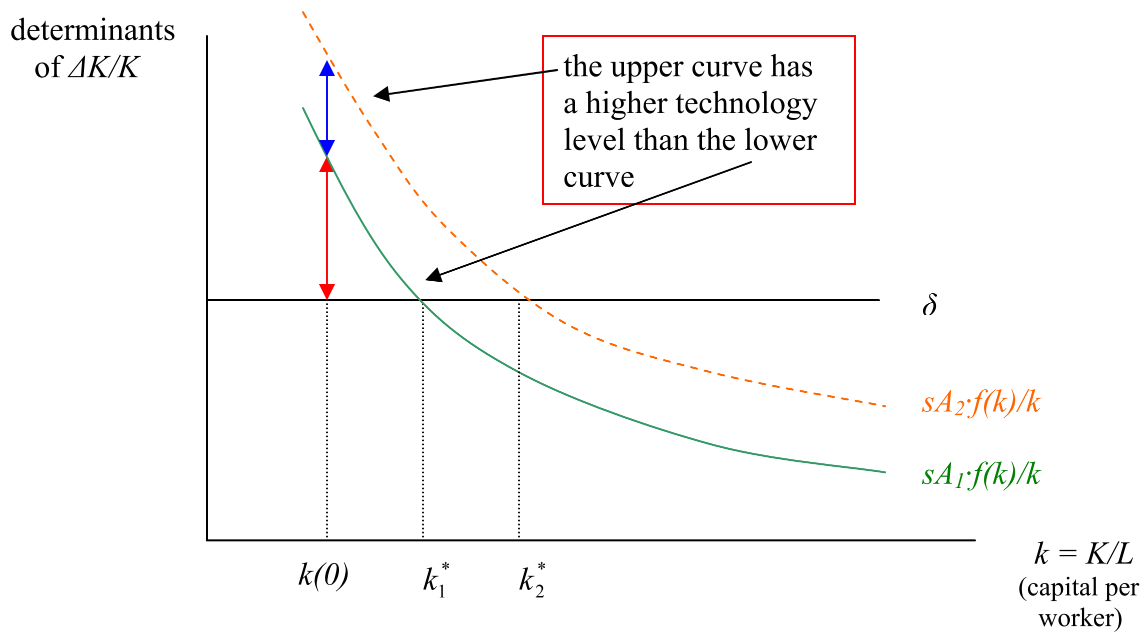


Figure 4.2

Effect of an Increase in the Technology Level in the Solow Model

This graph comes from Figure 3.9. The two curves for $s{\cdot}(y/k) = sA{\cdot}f(k)/k$ are for the technology levels A_1 and A_2, where A_2 is greater than A_1. The growth rate of capital, $\Delta K/K$, is higher at any k when the technology level is higher. For example, at *k(0),* when the technology level is A_1, $\Delta K/K$ equals the vertical distance shown by the lower red arrows. When the technology level is A_2, $\Delta K/K$ equals the vertical distance that adds in the upper blue arrows. In the steady state, $\Delta K/K$ is zero, regardless of the technology level. The higher technology level yields a higher steady-state capital per worker, that is, k_2^* is greater than k_1^*.

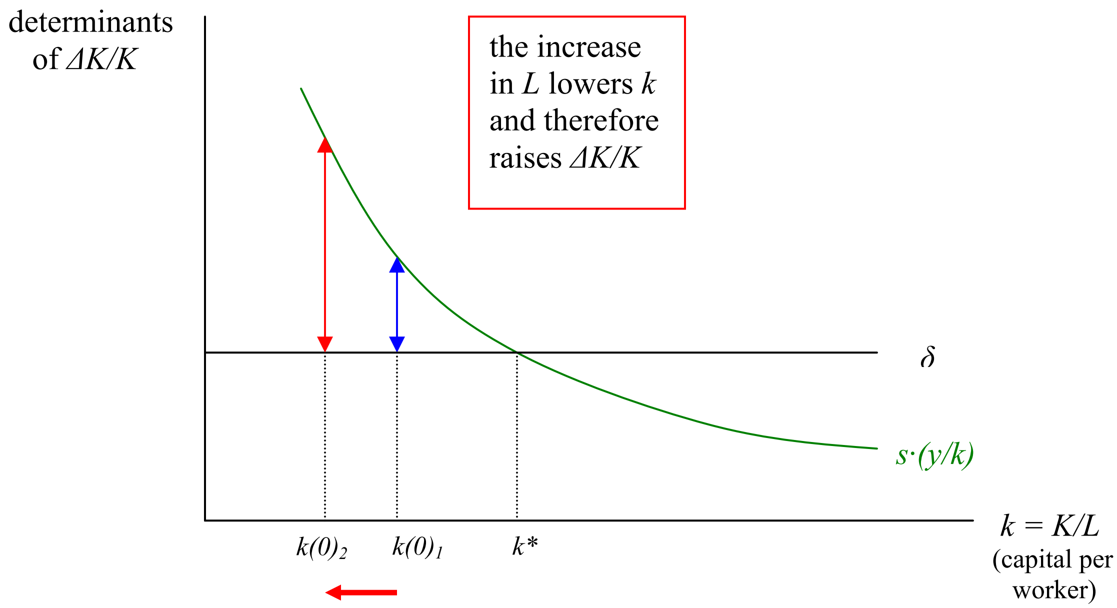


Figure 4.3

Effect of an Increase in Labor Input in the Solow Model

This graph comes from Figure 3.9. If labor input rises from L_1 to L_2, the initial capital per worker declines from $k(0)_1 = K(0)/L_1$ to $k(0)_2 = K(0)/L_2$. Therefore, the growth rate of capital, $\Delta K/K$, rises initially. Note that the vertical distance shown by the red arrows is larger than that shown by the blue arrows. The steady-state capital per worker, k^*, is the same for the two values of L.

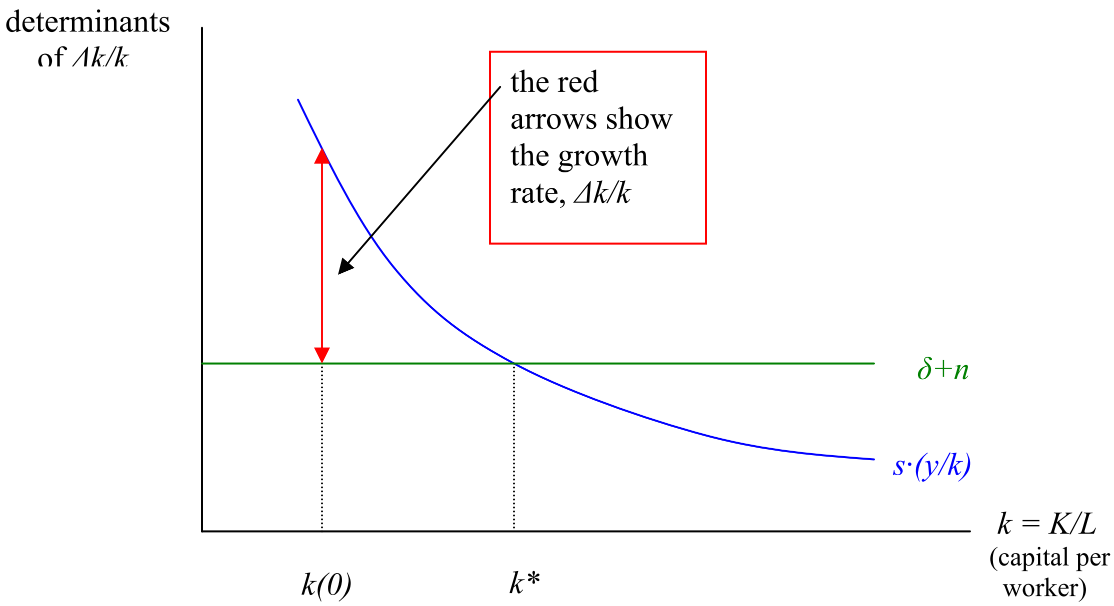


Figure 4.4

The Solow Model with Population Growth

This graph modifies Figure 3.9 to include population growth at the rate n. The effective depreciation line is at $\delta+n$, rather than δ. The vertical axis now shows determinants of the growth rate of capital per worker, $\Delta k/k$. This growth rate equals the vertical distance between the $s\cdot(y/k)$ curve and the effective depreciation line, $\delta+n$. At $k(0)$, $\Delta k/k$, shown by the red arrows, is greater than zero but then declines toward zero as k approaches its steady-state value, k^*.

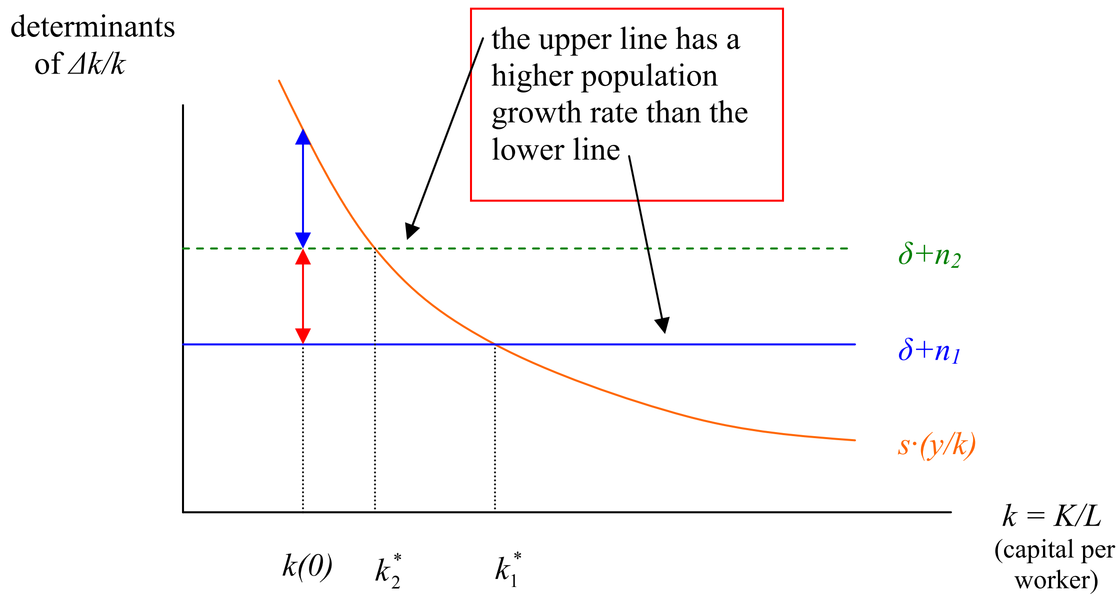


Figure 4.5

Effect of an Increase in the Population Growth Rate in the Solow Model

This graph comes from Figure 4.4. An increase in the population growth rate from n_1 to n_2 raises the line for effective depreciation from $\delta+n_1$ to $\delta+n_2$. The growth rate of capital per worker, $\Delta k/k$, is lower at any k when the population growth rate is higher. For example, at $k(0)$, when the population growth rate is n_1, $\Delta k/k$ equals the vertical distance given by the sum of the blue and red arrows. When the population growth rate is n_2, $\Delta k/k$ equals the vertical distance given just by the upper blue arrows. In the steady state, $\Delta k/k$ is zero, regardless of the population growth rate. A higher rate of population growth yields a lower steady-state capital per worker, that is, k_2^* is less than k_1^*.

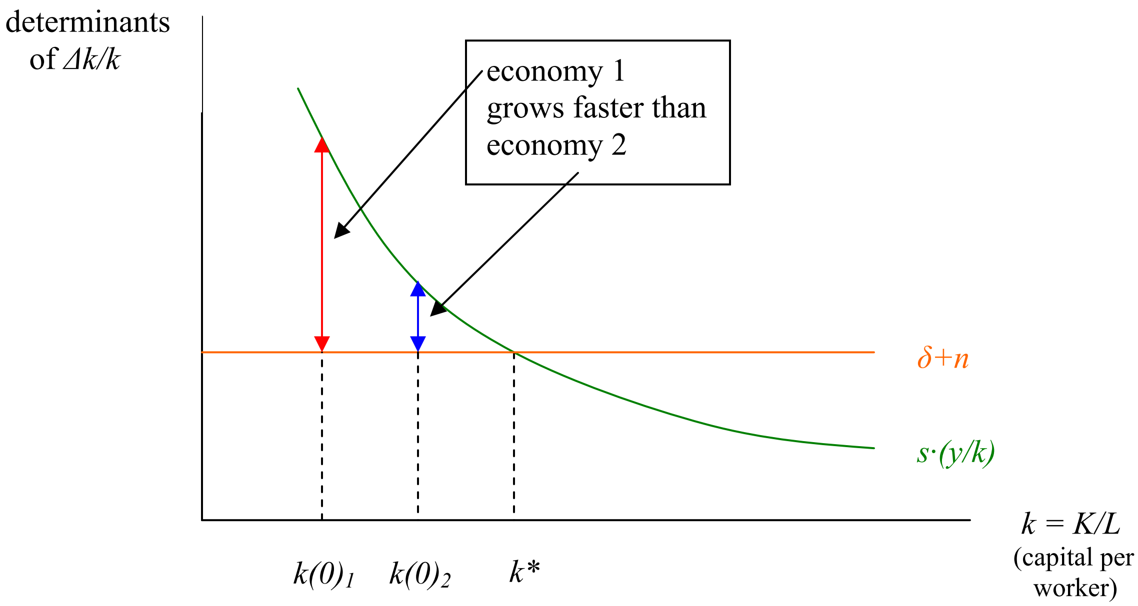


Figure 4.6

Convergence in the Solow Model

Economy 1 starts with lower capital per worker than economy 2—$k(0)_1$ is less than $k(0)_2$. Economy 1 grows faster initially because the vertical distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line is greater at $k(0)_1$ than at $k(0)_2$. That is, the distance marked by the red arrows is greater than that marked by the blue arrows. Therefore, capital per worker in economy 1, k_1, converges over time toward that in economy 2, k_2.

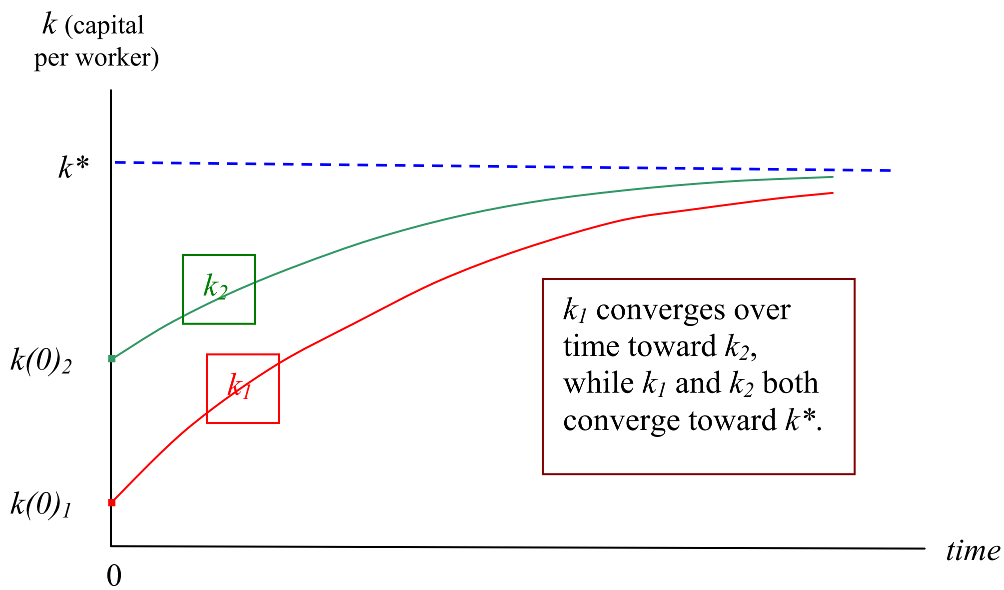


Figure 4.7

Convergence and Transition Paths for Two Economies

Economy 1 starts at the capital per worker $k(0)_1$ and economy 2 starts at $k(0)_2$, where $k(0)_1$ is less than $k(0)_2$. The two economies have the same steady-state capital per worker, k^*, shown by the dashed blue line. In each economy, k rises over time toward k^*. However, k grows faster in economy 1 because $k(0)_1$ is less than $k(0)_2$. (See Figure 4.6.) Therefore, k_1 converges over time toward k_2.

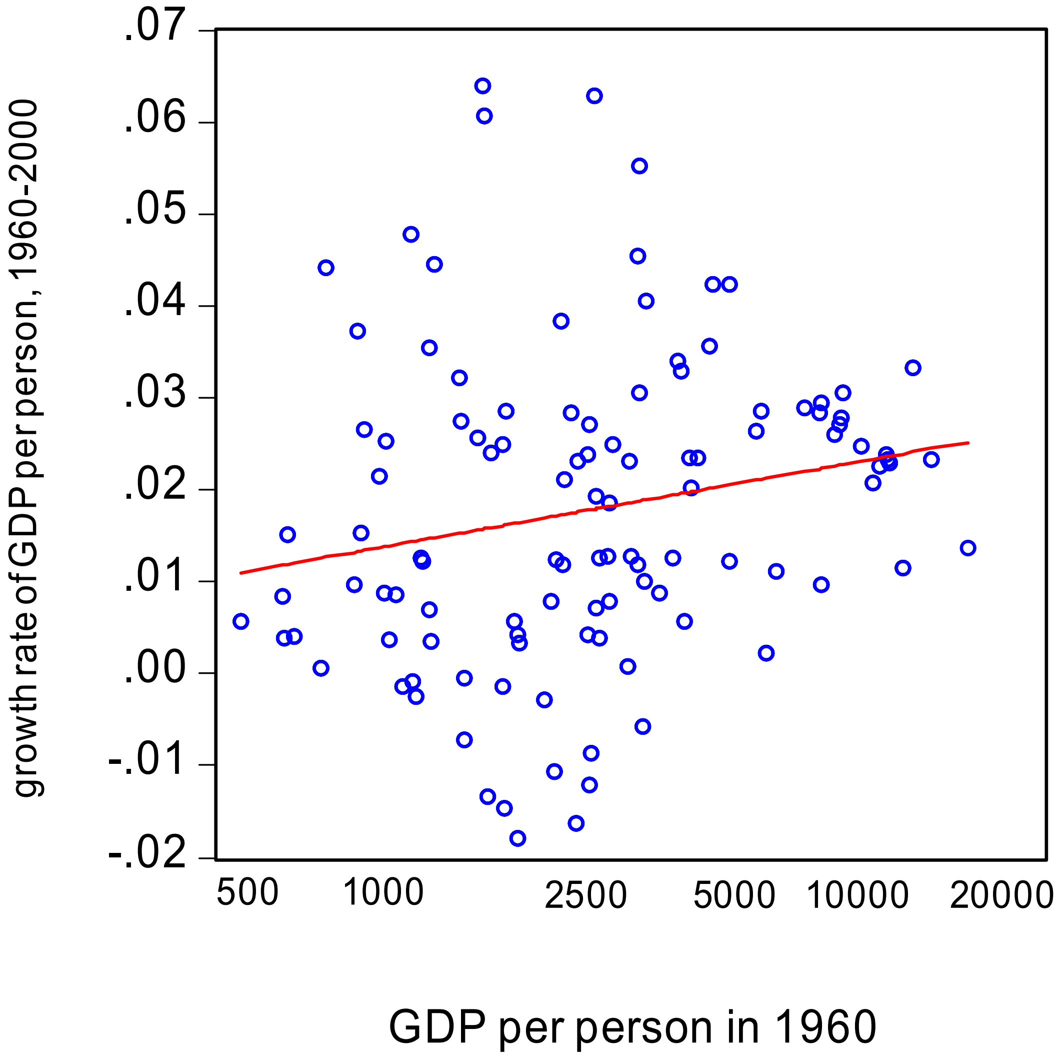


Figure 4.8

Growth Rate versus Level of Real GDP per Person for a Broad Group of Countries

The horizontal axis shows real GDP per person in 1960 in 1995 U.S. dollars on a logarithmic scale for 111 countries. The vertical axis shows the growth rate of real GDP per person for each country from 1960 to 2000. The red line is the straight line that provides a best fit to the relation between the growth rate of real GDP per person (the variable on the vertical axis) and the level of real GDP per person (on the horizontal axis). Although this line slopes upward, the slope is—in a statistical sense—negligibly different from zero. Hence, the growth rate is virtually unrelated to the level of real GDP per person. Thus, this broad group of countries does not display convergence.

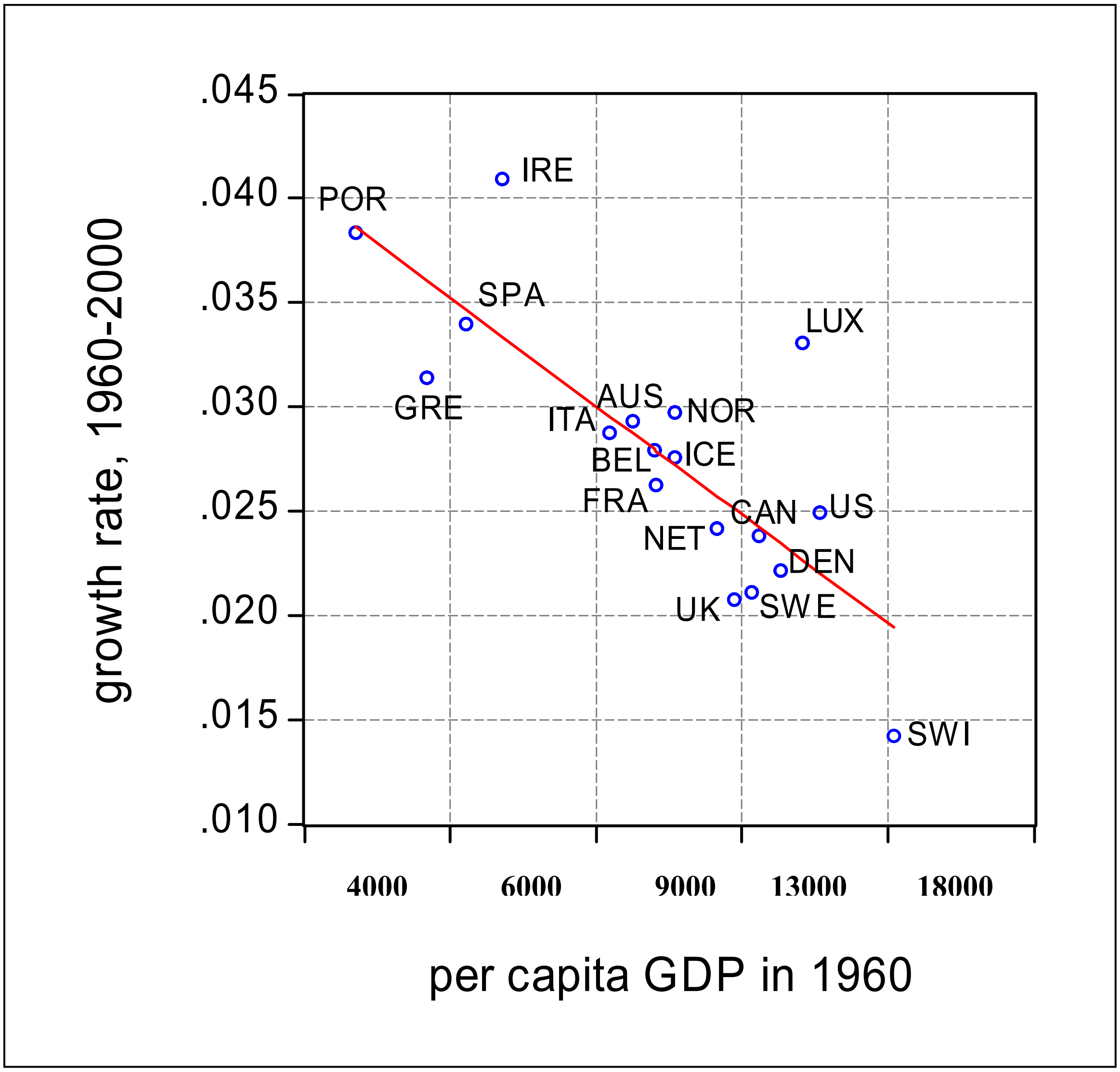


Figure 4.9
Growth Rate versus Level of Real GDP per Person for OECD Countries

The horizontal axis shows real GDP per person in 1960 in 1995 U.S. dollars on a logarithmic scale for 18 of the 20 founding members of the OECD (excluding Germany and Turkey). The abbreviation identifies each country. The vertical axis shows the growth rate of real GDP per person for each country from 1960 to 2000. The red line is the straight line that provides a best fit to the relation between the growth rate of real GDP per person (the variable on the vertical axis) and the level of real GDP per person (on the horizontal axis). The line has a clear negative slope—therefore, a lower level of real GDP per person in 1960 matches up with a higher growth rate of real GDP per person from 1960 to 2000. Thus, the OECD countries exhibit convergence.

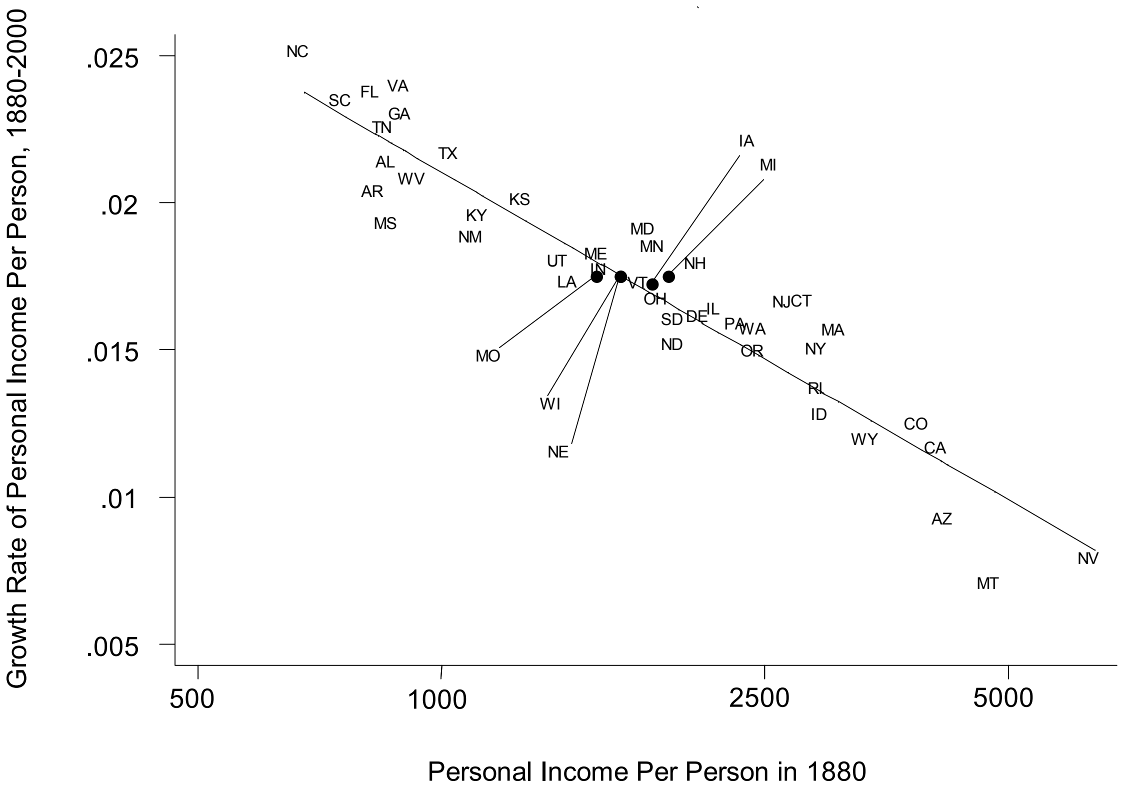


Figure 4.10

Growth Rate versus Level of Income per Person for U.S. States, 1880-2000

The horizontal axis shows real personal income per person in 1982-84 U.S. dollars on a logarithmic scale for 47 U.S. states. The two-letter abbreviation identifies the state. (Alaska, the District of Columbia, Hawaii, and Oklahoma are excluded.) The vertical axis shows the growth rate of real personal income per person for each state from 1880 to 2000. The solid line is the straight line that provides a best fit to the relation between the growth rate of income per person (the variable on the vertical axis) and the level of income per person (on the horizontal axis). The line has a clear negative slope—therefore, a lower level of income per person in 1880 matches up with a higher growth rate of income per person from 1880 to 2000. Thus, the U.S. states exhibit convergence.

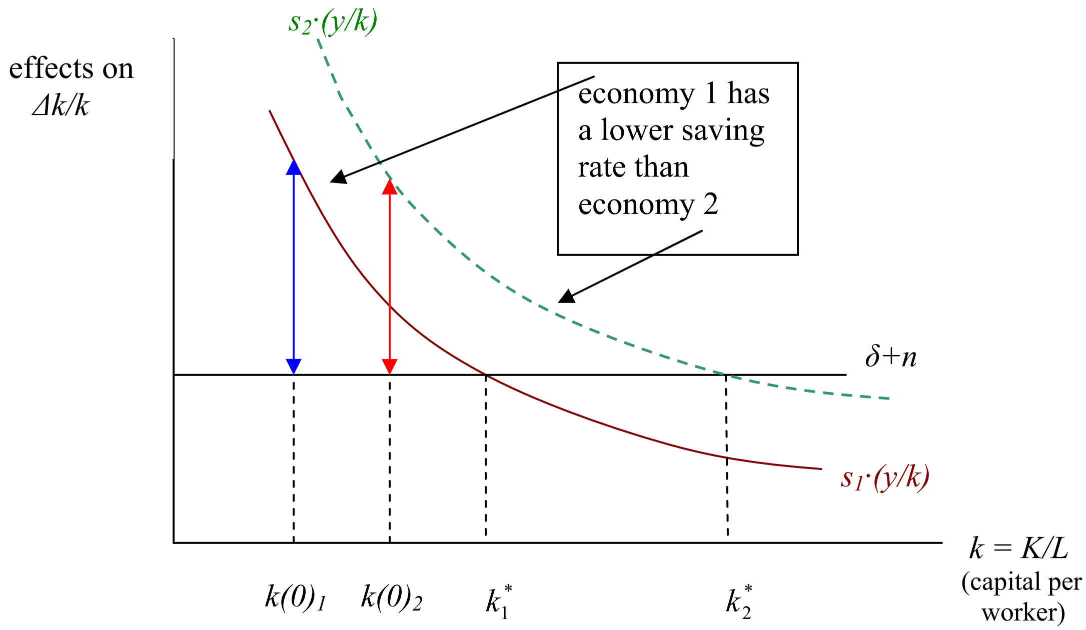


Figure 4.11

Failure of Convergence in the Solow Model:
Differences in Saving Rates

As in Figure 4.6, economy 1 starts with lower capital per worker than economy 2—$k(0)_1$ is less than $k(0)_2$. However, we now assume that economy 1 also has a lower saving rate, that is, s_1 is less than s_2. The two economies have the same technology levels, A, and population growth rates, n. Therefore, k_1^* is less than k_2^*. In this case, it is uncertain which economy grows faster initially. The vertical distance marked with the blue arrows may be larger or smaller than the one marked with the red arrows.

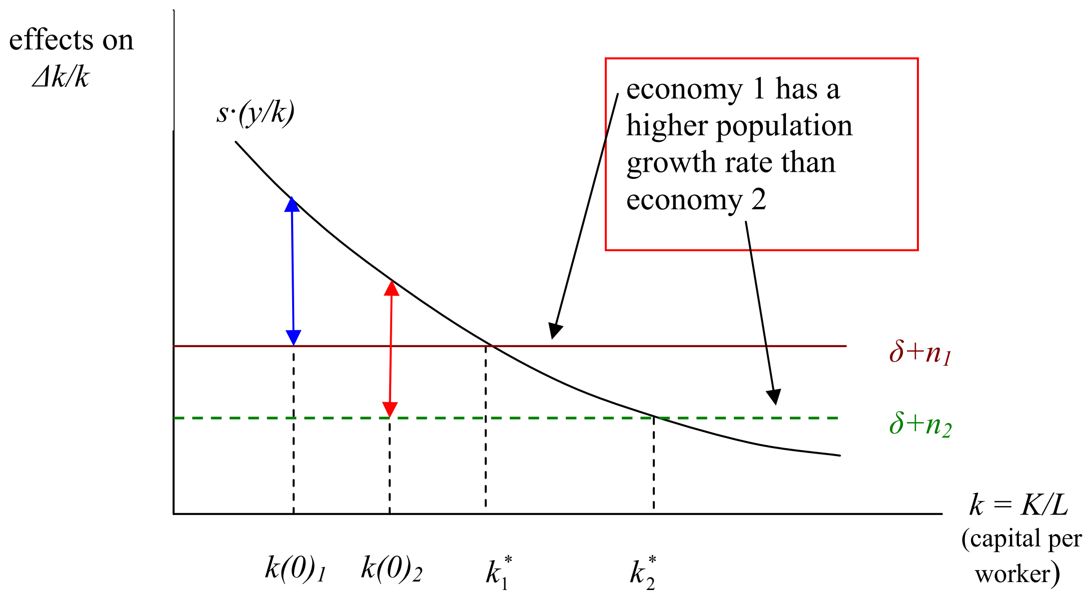


Figure 4.12

Failure of Convergence in the Solow Model: Differences in Population Growth Rates

As in Figure 4.11, economy 1 starts with lower capital per worker than economy 2—$k(0)_1$ is less than $k(0)_2$. The two economies now have the same saving rates, *s*, and technology levels, *A*, but economy 1 has a higher population growth rate, *n,* that is, n_1 is greater than n_2. Therefore, k_1^* is again less than k_2^*. As in Figure 4.11, it is uncertain which economy grows faster initially. The vertical distance marked with the blue arrows may be larger or smaller than the one marked with the red arrows.

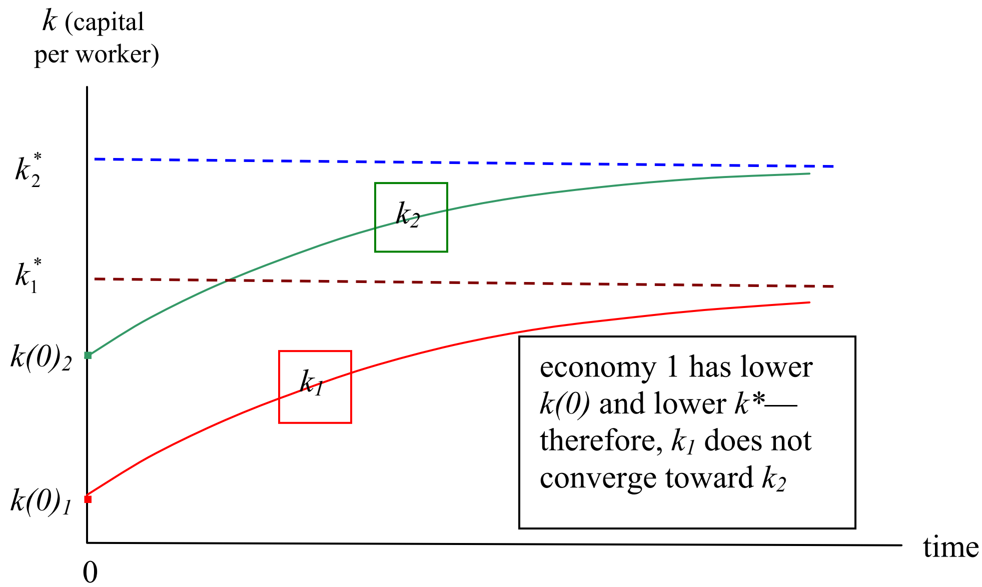


Figure 4.13

Failure of Convergence and Transition Paths for Two Economies

As in Figures 4.11 and 4.12, economy 1 has a lower starting capital per worker—$k(0)_1$ is less than $k(0)_2$—and also has a lower steady-state capital per worker—k_1^* (the dashed brown line) is less than k_2^* (the dashed blue line). Each capital per worker converges over time toward its own steady-state value: k_1 (the red curve) toward k_1^* and k_2 (the green curve) toward k_2^*. However, since k_1^* is less than k_2^*, k_1 does not converge toward k_2.

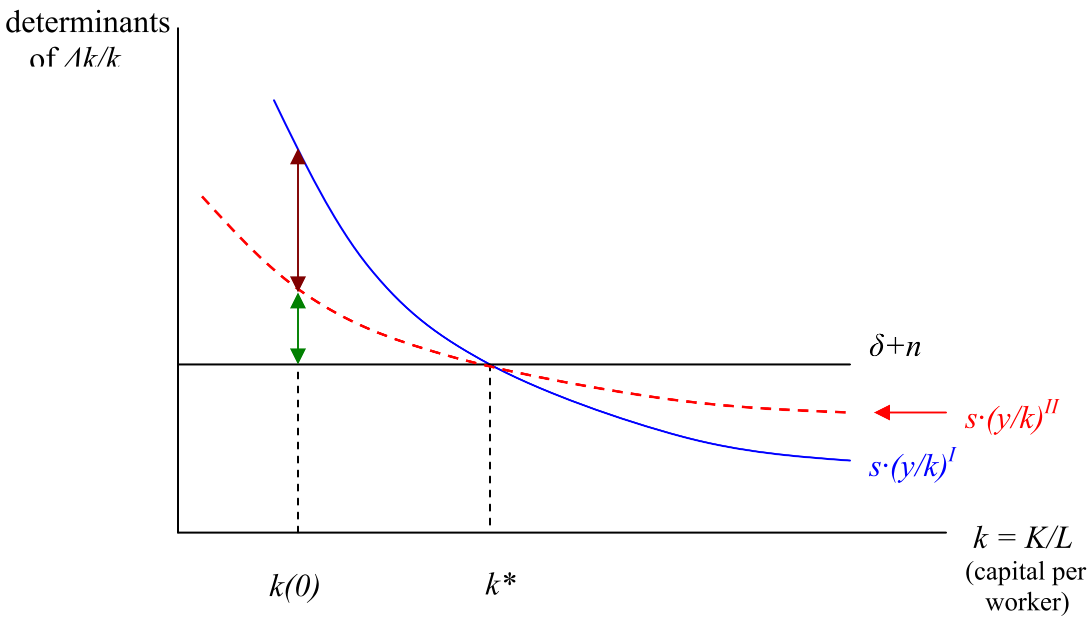


Figure 4.14

Determining the Speed of Convergence

This graph modifies the one in Figure 4.4. The first saving curve, $s\cdot(y/k)^{I}$, is the same as before. The second saving curve, $s\cdot(y/k)^{II}$, does not slope downward as much as the first one. The reason is that the average product of capital, y/k, diminishes less rapidly with k in the second case. At $k(0)$, the distance between the $s\cdot(y/k)$ curve and the $\delta+n$ line is greater in the first case (corresponding to the sum of the green and brown arrows) than the second (corresponding only to the green arrows). Therefore, the initial $\Delta k/k$ is higher in the first case, and the convergence to the steady state is faster. The conclusion is that convergence is faster when the average product of capital diminishes more rapidly with k.

January 12, 2005

Chapter 5

Conditional Convergence; Long-Run Economic Growth

In the previous two chapters, we developed and extended the Solow model of economic growth. For short-run analysis, the most important results were about convergence. We show in the first part of this chapter how these results can be used to understand many patterns of economic growth in the world.

We observed at the end of the last chapter that the major deficiency of the Solow model was its failure to explain long-run economic growth. In the second part of this chapter, we extend the model to allow for long-run growth.

I. Conditional Convergence in Practice

We found in the Solow model that the steady-state capital per worker, k^*, was an important determinant of economic growth. We summarized this conclusion in an equation:

$$\Delta k/k = \varphi[\underset{(-)}{k(0)}, \underset{(+)}{k^*}]. \qquad (4.13)$$

Thus, for a given starting capital per worker, $k(0)$, an increase in k^* raised the growth rate, $\Delta k/k$.

In our previous discussion, we focused on three variables that influenced k^*—the saving rate, s, the technology level, A, and the population growth rate, n. Recent research has extended the Solow model to allow for additional variables that affect the steady-state capital and real GDP per worker, k^* and y^*. We can understand these effects, without working through the details, by taking a broader view of the technology level, A. The important thing about a higher A is that it raises productivity, that is, it allows real GDP to rise for given inputs of capital and labor. Many variables that are not strictly technological also influence an economy's productivity. These other influences affect economic growth in ways analogous to changes in A.

As an example, productivity depends on the degree of market efficiency. Economies can enhance efficiency by removing restrictions due to government regulations, by lowering tax rates, and by promoting competition, possibly through anti-trust enforcement. Another way for markets to work better is for governments to allow free trade in goods and services across international borders. This kind of international openness allows countries to specialize in the production of the goods and services in which they have natural advantages. Hence, greater international openness tends to raise productivity. An additional influence on productivity concerns legal and political systems. Productivity tends to rise if governments do better at maintaining private property rights, if the judicial system runs more smoothly, and if official corruption declines.

A. Recent research on the determinants of economic growth

Recent research has used the equation for conditional convergence, equation (4.13), as a framework to analyze the determinants of economic growth across countries. The idea is to measure an array of variables, each of which influences a country's steady-state capital per worker, k^*. Equation (4.13) then tells us two things.

First, if we hold fixed k^* (by holding fixed the variables that influence k^*), the growth rate of capital per worker, $\Delta k/k$, should exhibit convergence. That is, for given k^*, a lower $k(0)$ should match up with a higher $\Delta k/k$. Second, any variable that raises or lowers k^* should correspondingly raise or lower $\Delta k/k$ for given $k(0)$.

Figure 5.1 shows empirical results for the relation between the growth rate and level of real GDP per person. The cross-country data are for a broad group of countries and are essentially the same as those plotted in Figure 4.8.[1] However, because we are holding fixed variables that determine the steady-state position, k^*, the graph looks very different from before. With the other variables held constant, the convergence pattern becomes clear—low levels of real GDP per person match up with high growth rates, and high levels of real GDP per person match up with low growth rates. Thus, there is a lot of evidence for conditional convergence across a broad group of countries.

The relation shown in Figure 5.1 applies when we hold constant a list of variables that influence k^*. The particular list used to construct the graph is

- a measure of the saving rate;
- the fertility rate for the typical woman (which influences population growth);
- subjective measures of maintenance of the rule of law and democracy;
- the size of government, as gauged by the share of government consumption expenditures in GDP;

[1] The one new feature is that the data are for the three ten-year periods from 1965 to 1995. A country's level of real GDP per person in 1965 is matched with its growth rate of real GDP per person from 1965 to 1975, the level of real GDP per person in 1975 is matched with the growth rate of real GDP per person from 1975 to 1985, and so on. If the data are available, each country appears three times in the graph. In contrast, in Figure 4.8, a country's level of real GDP per person in 1960 is matched with its growth rate of real GDP per person from 1960 to 2000. Therefore, each country appears only once in this graph.

- the extent of international openness, measured by the volume of exports and imports;
- changes in the terms of trade, which is the ratio of prices of exported goods to prices of imported goods;
- measures of investment in education and health; and
- the average rate of inflation, which is an indicator of macroeconomic policy.

One reason that we considered these variables is to isolate conditional convergence, as shown in Figure 5.1. Equally important, however, is that we learn how the variables in the list affect economic growth. The research shows that the growth rate of real GDP per person rises in response to a higher saving rate, lower fertility, better maintenance of the rule of law, smaller government consumption, greater international openness, improvement in the terms of trade, qreater quantity and quality of education, better health, and lower inflation. Democracy has a less clear effect—if a country starts from a totalitarian system, increases in democracy seem to favor economic growth. However, after a country reaches a mid-range of democracy (characteristic in recent years of Indonesia, Turkey, and several places in Latin America), further democratization seems to reduce growth.

Research on the determinants of economic growth has been lively since the early 1990s. This research has suggested numerous additional variables that influence growth. The variables considered include the scope of financial markets, the degree of income inequality, the extent of official corruption, the role of colonial and legal origins, and the intensity of religious participation and beliefs. In the nearby box, we consider two other variables, debt relief and foreign aid.

These kinds of empirical results have raised our knowledge about the determinants of economic growth, but our understanding remains incomplete. For one thing, economists have isolated only some of the variables that influence growth. The problems relate partly to data—for example, it is difficult to quantify government distortions from regulations and taxation or to measure various aspects of legal and political systems. Another problem is that many variables influence economic growth, and it is impossible to isolate all of these effects with the limited data available. Moreover, it is often difficult to be sure whether a variable—for example, maintenance of the rule of law or the levels of investment in education and health—affects economic growth or is affected by growth. In practice, both directions of causation are often important.

A rock star's perspective on debt relief and foreign aid

In the summer of 1999, I met Bono, the lead singer of the rock group U2. Bono wanted to discuss the Jubilee 2000 campaign, which was a global movement aimed at canceling the international debts of the world's poorest countries. I told him that I was an unlikely candidate to support Jubilee 2000. Bono said that was precisely why he wanted to talk with me. He wanted to see whether a hard-thinking economist could be persuaded of the soundness of the campaign.[2] In particular, he was not interested in a global welfare project but rather wanted to push debt relief as a way to promote sound economic policies and high economic growth. He even said that the relief would be conditioned on a country's commitment to use the freed-up money for productive investments.

I was shocked to hear these arguments from a rock star. Nevertheless, I recovered to say that this commitment would be unenforceable and that debt relief would not be on the top-ten list of policies for growth promotion in poor countries. More important were well-functioning legal and political systems, openness to markets at home and abroad, investments in education and health, and sound macroeconomic policy. I mentioned the musical line "money for nothing" (from a song by Dire Straits) and said that it applied to a number of ways in which a country received unearned money. These included debt relief, debt default, foreign aid, and even natural resources such as oil. Experience showed that all of these forms of free money tended to reduce economic growth. I also argued that growth would be encouraged if a country gained a reputation for honoring foreign debts and other contracts.

Bono agreed that it was important for a country to fulfill its debt obligations, especially those that originated from sensible commercial transactions. However, he argued that most of the international debts of African

[2] The "hard-thinking" phrase was Bono's. I think he meant that I was a conservative economist who would tend to favor free markets and oppose large government. (Personally, I prefer the phrases classical liberal and libertarian to conservative.)

and other poor countries derived from poorly designed projects conceived by the World Bank, other international organizations, and donor countries such as the United States. Many of these loans were made to corrupt dictators, who diverted the funds for personal gain. He noted that these debts could never be repaid. Bono said that the idea of the term Jubilee 2000 was that it was a one-time happening and would, therefore, not encourage default on newly incurred debts. (I was a little worried here, because the Bible says that Jubilees are supposed to occur every 50 years.)

In the end, I was not convinced to put debt relief on the top-ten list of growth-promoting policies for poor countries. But, since the arguments I heard were better than I had anticipated, I was pleased at the time to offer two restrained cheers for Jubilee 2000.

In retrospect, my response was two cheers too many. The economist Bill Easterly has argued convincingly (in *The Elusive Quest for Growth*, MIT Press, 2001) that the problem of high foreign debt for poor countries is not new and that the remedy of debt relief is neither new nor effective. Easterly noted that we have already been trying debt forgiveness for two decades, with little of the salutary results promised by Jubilee 2000. He also demonstrated that the main response historically to debt relief has been for countries to run up new debts, most of which were used to finance non-productive projects. There is no evidence that past debt relief operations helped the poor, which were the intended target of Bono. So, why would one expect new debt relief to work any better?

Although I had doubts about the efficacy of Bono's proposals, he nevertheless achieved many successes since our meeting in summer 1999. His campaign brought him into contact with numerous world leaders, including then President Bill Clinton and the pope (who is said to have tried on Bono's famous sunglasses). Bono swayed numerous politicians and economists to his cause, and his great exercise in persuasion culminated in the $435 million U.S. debt relief legislation of November 2000.

Because I hold Bono in high esteem, I wish I could believe that this and future programs of debt relief would help to spur economic growth. But my understanding of economics and my research keep me from believing these things. I wonder what would happen if Bono, instead, directed his persuasive talents to furthering ideas that seem to matter for economic growth. I have in mind property rights, the rule of law, free markets, and small government. And, I would be happy to include investments in education and health. But, of course, this is just a dream.

B. Examples of conditional convergence

If we look at history, we can find many examples of conditional convergence. At the end of World War II, the economies of many nations were destroyed. Cities were leveled, factories bombed, and farmland used as battlefields. By 1946, Japan, Germany, France, and other countries in Europe suffered sharp reductions in physical capital, which resulted in a low starting value of capital per worker, $k(0)$.[3] But these countries also had characteristics that were favorable to rapid economic recovery—including strong human capital in the forms of education and health and good legal and political traditions that encouraged markets and trade. We can represent these favorable characteristics as a high value of the steady-state capital per worker, k^*. Hence, conditional convergence—as summarized in equation (4.13)—predicts that these countries would grow rapidly in the aftermath of World War II. This prediction fits with the facts.

As another example of conditional convergence, in the 1960s, many East Asian countries, such as South Korea and Taiwan, were poor and, therefore, had low values of $k(0)$. However, these countries also had reasonably good legal systems, strong programs in education and health, high saving rates, and relatively high openness to international trade. Therefore, k^* was high. Hence, we predict the high rates of economic growth that occurred from 1960 to 2000.

The typical sub-Saharan African country was also poor in the 1960s, that is, $k(0)$ was low. Hence, from the perspective of absolute convergence, we would predict high growth rates in Africa—whereas, in fact, the growth rates were the lowest in the world from 1960 to 2000. Conditional convergence can explain this outcome, because the

---

[3] These countries also experienced substantial losses in human capital, but the reductions in physical capital were quantitatively more significant.

African countries had poorly functioning legal and political systems, poor education and health programs, high rates of population growth, and large corrupt governments. Thus, k^* was low, and the sub-Saharan African countries failed to grow.

We see from these examples that the idea of conditional convergence allows us to understand many apparently dissimilar experiences about economic growth. This idea helps us to understand growth of rich countries after World War II, as well as of East Asian and sub-Saharan African countries from 1960 to 2000. More broadly, the idea of conditional convergence helps us to understand the range of growth experienced by a broad group of countries since 1960.

II. Long-Run Economic Growth

Thus far, the Solow model does not explain how capital and real GDP per worker, k and y, grow in the long run. In the model, these variables are constant in the long run at their steady-state values, k^* and y^*. Thus, the model does not explain how real GDP per person grew at around 2% per year for well over a century in the United States and other rich countries.

We now consider extensions of the Solow model that explain long-run economic growth. We start with a model in which the average product of capital, y/k, does not diminish as capital per worker, k, rises. Then we allow for technological progress in the sense of continuing growth of the technology level, A. We consider first a model in which this technological progress is just assumed—that is, A grows in an exogenous manner. Then we consider theories in which technological progress is explained within the model—that is, endogenous growth models. We also consider models of technological diffusion, in which a country's technology level, A, rises through imitation of advanced technologies from other countries.

A. Models with constant average product of capital

The diminishing average product of capital, y/k, plays a major role in the transition phase of the Solow growth model. As capital per worker, k, increases, the decline in y/k reduces the growth rate, $\Delta k/k$. Eventually, the economy approaches a steady state, in which k is constant at k^* and $\Delta k/k$ is zero. This sketch of the transition suggests that the conclusions would be very different if y/k did not decline as k rose. Thus, we now consider a modified model in which y/k is constant. We are particularly interested in whether this modification can explain economic growth in the long run.

Recall that the growth rate of capital per worker, k, is given by

$$\Delta k/k = s\cdot(y/k) - (\delta+n). \tag{4.8}$$

Now we want to reconsider our assumption that the average product of capital, y/k, falls as k rises. The diminishing average product makes sense if we interpret capital narrowly, for example, as machines and buildings. If a business keeps expanding its machines and buildings, without adding any workers, we would expect the marginal and average products of capital to fall. In fact, if labor input did not increase, we would expect the marginal product of capital eventually to get close to zero. If no one is available to operate an extra machine, the marginal product of that machine would be nil.

Another argument, however, is that we should interpret capital broadly, especially to include human capital in the forms of formal education, on-the-job training, and health. The idea is that human capital is productive and that the amount of this capital can be increased by investment. Hence, human capital is analogous to machines and buildings. We might also go further to include forms of capital that are often owned by government, such as **infrastructure capital** used for transportation and utilities.

The tendency for capital's marginal and average products to fall as k rises is less pronounced and may even be absent if we view capital in this broad sense. That is, if we double not only machines and buildings but also human and infrastructure capital, real GDP may roughly double. The only thing we are holding constant here, aside from the technology level, is the quantity of raw labor. If raw labor is not a critical input to production, capital's marginal and average products may not decline as capital accumulates.

To see the consequences of this modification, consider a case in which capital—broadly defined to include human and infrastructure capital—is the only factor input to production. Then, instead of the usual production function,

(3.2) $$y = A \cdot f(k),$$

we might have

(5.1) $$y = Ak.$$

Equation (5.1) is the special case of equation (3.2) in which $f(k) = k$. For obvious reasons, the new model is called the ***Ak* model**.

In the Ak model, the average product of capital is constant. If we divide both sides of equation (5.1) by k, we get

(5.2) $$y/k = A.$$

That is, the average product of capital equals the technology level, A.[4] If we substitute $y/k = A$ into equation (4.8), we get that the growth rate of k is

[4] The marginal product of capital is also constant and equal to A.

(5.3) $$\Delta k/k = sA - (\delta+n).$$

We can use a graph analogous to Figure 4.4 to study the determination of the growth rate, $\Delta k/k$, in the Ak model. The difference now, in Figure 5.2, is that the term $s\cdot(y/k) = sA$ is not downward sloping versus k—instead, it is a horizontal line at sA. The effective depreciation line at $\delta+n$ is the same as before. Also as before, the growth rate, $\Delta k/k$, equals the vertical distance between the two lines. However, now this distance is constant, rather than diminishing as k rises.

Two important conclusions follow from Figure 5.2. First, instead of being zero, the long-run growth rate, $\Delta k/k$, is greater than zero and equal to $sA - (\delta+n)$, as shown in the graph and in equation (5.3). This growth rate is greater than zero because we assumed that sA was greater than $\delta+n$. This condition is more likely to hold the higher are the saving rate, s, and the technology level, A, and the lower is the effective depreciation rate, $\delta+n$.

If sA is greater than $\delta+n$, as assumed in Figure 5.2, growth of capital per worker, k, continues forever at the rate $sA - (\delta+n)$. Moreover, since $y = Ak$, real GDP per worker, y, grows forever at the same rate. In this case, a higher saving rate, s, or a higher technology level, A, raises the long-run growth rates of capital and real GDP per worker, whereas a higher effective depreciation rate, $\delta+n$, lowers the long-run growth rates. In contrast, in the standard Solow model, the long-run growth rates of capital and real GDP per worker were zero and, therefore, did not depend on s, A, δ, and n. The reason for the different result is that the standard model assumed diminishing average product of capital.

The second important result from Figure 5.2 and equation (5.3) is the absence of convergence. The growth rates of capital and real GDP per worker do not change as k and y rise.

Economists have developed more sophisticated models in which the average product of capital remains constant as capital accumulates. Some models distinguish human from non-human capital and allow for an education sector that produces human capital. However, two basic shortcomings apply to most of these models. First, the loss of the convergence prediction is a problem, because conditional convergence is an important empirical phenomenon. Therefore, we cannot be satisfied with a growth model that fails to predict conditional convergence. Second, a common view among economists is that diminishing marginal and average products of capital apply eventually to the accumulation of capital even when interpreted in a broad sense to include human and infrastructure capital. If we reintroduced diminishing average product of capital, growth could not continue in the long run just by accumulating capital. Therefore, we now turn to another explanation for long-run economic growth: technological progress.

B. Exogenous technological progress

In the previous chapter, we studied the effects of a one-time increase in the technology level, A. This change raised the growth rates of capital and real GDP per worker, $\Delta k/k$ and $\Delta y/y$, during the transition to the steady state. However, the economy still approached a steady state in which the growth rates were zero. Thus, we cannot explain long-run economic growth from a single increase in A. Rather, we have to allow for continuing increases in A. This regular process of improvements in the technology is called **technological progress**.

Solow did extend his growth model to allow for technological progress, but he did not try to explain the sources of this progress. He just assumed that technological progress occurred and then examined the consequences for economic growth. In other words, he treated technological progress as exogenous—coming from outside of the model. This approach would be reasonable if most improvements in technology come by luck, in particular, if they do not depend much on purposeful effort by businesses and workers. In this section, we follow Solow's practice by assuming that the technology level, A, grows exogenously at a constant rate g:

$$\Delta A/A = g.$$

In a later section, we discuss **endogenous growth theories**, which try to explain the rate of technological progress within the model.

1. The steady-state growth rate. The growth-accounting equation is again

$$\Delta Y/Y = \Delta A/A + \alpha\cdot(\Delta K/K) + (1\text{-}\alpha)\cdot(\Delta L/L). \tag{3.3}$$

If we substitute $\Delta A/A = g$ and $\Delta L/L = n$, the population growth rate, we get

$$\Delta Y/Y = g + \alpha\cdot(\Delta K/K) + (1\text{-}\alpha)\cdot n. \tag{5.4}$$

Recall that the growth rate of real GDP per worker, $\Delta y/y$, is given by

$$\Delta y/y = \Delta Y/Y - n. \tag{4.6}$$

If we substitute for $\Delta Y/Y$ from equation (5.4), we get

$$\Delta y/y = g + \alpha\cdot(\Delta K/K) + (1\text{-}\alpha)\cdot n \text{ - } n$$

$$= g + \alpha\cdot(\Delta K/K) + \cancel{n} - \alpha n - \cancel{n}$$

$$= g + \alpha\cdot(\Delta K/K - n).$$

Since the growth rate of capital per worker, $\Delta k/k$, is given by

(4.5) $$\Delta k/k = \Delta K/K - n,$$

we can substitute $\Delta k/k$ for $\Delta K/K - n$ to get

(5.5) $$\Delta y/y = g + \alpha\cdot(\Delta k/k).$$

We therefore see that real GDP per worker grows because of technological progress, g, and growth of capital per worker, $\Delta k/k$.

The growth rate of capital per worker, $\Delta k/k$, is still determined from

(4.14) $$\Delta k/k = sA\cdot f(k)/k - (\delta+n).$$

If we substitute this expression for $\Delta k/k$ into equation (5.5), we get

(5.6) $$\Delta y/y = g + \alpha\cdot[sA\cdot f(k)/k - (\delta+n)].$$

In our previous analysis, where A was constant, increases in k led to reductions in the average product of capital, $y/k = A\cdot f(k)/k$. Consequently, in the long run, the economy reached a steady state in which the average product of capital was low enough so that the growth rate $\Delta k/k$ was zero in equation (4.14). Then, with $g = 0$, the growth rate of real GDP per worker, $\Delta y/y$, also equaled zero in equations (5.5) and (5.6).

The difference now is that each increase in A raises the average product of capital, $y/k = A\cdot f(k)/k$, for given k. Hence, the negative effect of rising k on capital's average product, y/k, is offset by a positive effect from rising A. The economy will tend toward a situation in which these two forces balance. That is, k will increase in the long run at a constant rate, and the average product of capital, y/k, will be constant. We call this situation **steady-state growth**.

Since the average product of capital, y/k, is constant during steady-state growth, the numerator of the ratio, y, must grow at the same rate as the denominator, k. Therefore, we have

(5.7) $(\Delta y/y)^* = (\Delta k/k)^*,$

where the asterisks designate values associated with steady-state growth.

We know from equation (5.7) that capital and real GDP per worker, k and y, grow at the same rate in steady-state growth. Now we want to determine the rate at which k and y are growing. To do this, we apply equation (5.5) to a situation of steady-state growth to get

(5.8) $(\Delta y/y)^* = g + \alpha\cdot(\Delta k/k)^*.$

Using equation (5.7), we can replace $(\Delta k/k)^*$ on the right-hand side by $(\Delta y/y)^*$ to get

$$(\Delta y/y)^* = g + \alpha\cdot(\Delta y/y)^*.$$

If we move the term $\alpha\cdot(\Delta y/y)^*$ from the right side to the left side, we get

$$(\Delta y/y)^* - \alpha\cdot(\Delta y/y)^* = g,$$

which implies, after we combine terms on the left,

$$(1-\alpha)\cdot (\Delta y/y)^* = g.$$

Finally, if we divide both sides by $1-\alpha$, we get the solution for the steady-state growth rate of real GDP per worker:

Key equation (steady-stage growth rate with technological progress):

(5.9) $(\Delta y/y)^* = g/(1-\alpha).$

Since $0<\alpha<1$, equation (5.9) tells us that the steady-state growth rate of real GDP per worker, $(\Delta y/y)^*$, is greater than the rate of technological progress, g:

the steady-state growth rate of real GDP per worker, (Δy/y)*,

is greater than the rate of technological progress, g.

As an example, if $\alpha = 1/2$, we have

$$(\Delta y/y)^* = 2g.$$

Thus, when $\alpha = ½$, $(\Delta y/y)^*$ is twice the rate of technological progress, g. For example, if $g = 1\%$ per year, $(\Delta y/y)^* = 2\%$ per year.

The reason that $(\Delta y/y)^*$ is greater than g is that the steady-state growth rate of capital per worker, $(\Delta k/k)^*$, is greater than zero, and this growth rate adds to g to determine $(\Delta y/y)^*$—see equation (5.8). In fact, we know from equation (5.7) that

$$(\Delta k/k)^* = (\Delta y/y)^*.$$

Therefore, equation (5.9) implies

$$(\Delta k/k)^* = g/(1\text{-}\alpha). \tag{5.10}$$

The important finding from equations (5.9) and (5.10) is that exogenous technological progress at the rate $\Delta A/A = g$ allows for long-term growth in real GDP and capital per worker at the rate $g/(1\text{-}\alpha)$. The technological progress offsets the tendency for the average product of capital, y/k, to fall when k rises and, thereby, allows for long-term growth in real GDP and capital per worker.

Recall from our discussion in chapter 3 that the growth rate of real GDP per person in the United States averaged 2.0% per year from 1869 to 2003. Similar long-term growth rates of real GDP per person—around 2% per year—applied to other advanced economies. To explain this long-term growth within the Solow model, we have to look at the model's predictions for steady-state growth.

Since the labor-force participation rate is constant in the model, the growth rate of real GDP per person equals the growth rate of real GDP per worker. Therefore, to get long-term growth of real GDP per person at around 2% per year, we need the steady-state growth rate of real GDP per worker, which equals $g/(1\text{-}\alpha)$ from equation (5.9), to be around 2% per year. If we think of α as the share of capital income and use values for α between 1/3 and 1/2, the required value for g is a little over 1% per year. In other words, if the technology improves exogenously at a rate around 1% per year, the Solow model's prediction for the long-term growth rate of real GDP per person matches the long-term growth rates observed in the United States and other advanced countries.

2. Steady-state saving. Now we consider how technological progress affects the amount saved in the steady state. The growth rate of capital per worker, $\Delta k/k$, is again

(4.8) $$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

We also have that, in a situation of steady-state growth, $\Delta k/k$ is given by

(5.10) $$(\Delta k/k)^* = g/(1-\alpha).$$

Therefore, in steady-state growth, we can replace $\Delta k/k$ in equation (4.8) by $g/(1-\alpha)$ to get

$$g/(1-\alpha) = s\cdot(y/k) - (\delta+n).$$

We can then rearrange the terms to get

$$s\cdot(y/k) = \delta+n + g/(1-\alpha).$$

If we multiply through by k, we get an equation that determines saving per worker, sy, in a situation of steady-state growth:

(5.11) $$sy = \delta k + nk + [g/(1-\alpha)]\cdot k \quad \textit{(in steady-state growth)}.$$

We considered before the determinants of saving per worker in the steady state when $g = 0$. In that case, in the steady state, saving covered two things, corresponding to the first two terms on the right-hand side of equation (5.11). The first term, δk, is the saving per worker needed to cover depreciation of capital, and the second term, nk, is the saving per worker required to provide the growing labor force with capital to work with.

Now, when g is greater than zero, we also have the third term, $[g/(1-\alpha)]\cdot k$. Since equation (5.10) tells us that $g/(1-\alpha)$ equals the steady-state growth rate of capital per worker, $\Delta k/k$, this term is

$$
\begin{aligned}
[g/(1-\alpha)]\cdot k &= (\Delta k/\cancel{k})\cdot\cancel{k} \\
&= \Delta k.
\end{aligned}
$$

Therefore, this term measures the saving per worker needed in the steady state to increase the capital per worker by the amount Δk.

3. The transition path and convergence. In Figure 3.10, we analyzed the transition path for capital per worker, k, in the model without technological progress. We found that k gradually approached its steady-state value, k^*. Thus, k^* was the target that k was approaching. The model with exogenous technological progress still has a transition path for k. However, we have to think of a moving target, rather than a fixed one.

In steady-state growth, equation (5.10) says that capital per worker rises at the rate $(\Delta k/k)^* = g/(1-\alpha)$. Hence, the level of capital per worker, k, is not constant in the steady state—k grows at the rate $g/(1-\alpha)$. In this situation, we can still define k^* to be the value that k takes at a point in time during steady-state growth. We just have to remember that k^* will not be constant over time when g is greater than zero.

Capital per worker, k, again starts at some initial value, $k(0)$. The model still has a transition in which k moves from $k(0)$ to its steady-state path. However, we have to represent the steady-state path not by a constant, but rather by the blue line labeled k^* in

Figure 5.3.[5] This line has a positive slope because capital per worker grows in the steady state. The graph shows that k begins at $k(0)$, rises over time along the red curve, and gradually approaches the moving target k^*.

Recall that, along the steady-state path, $k = k^*$ grows at the rate $g/(1-\alpha)$ (equation [5.10]). Therefore, in order for k to approach k^*, as shown in Figure 5.3, k must grow at a rate higher than $g/(1-\alpha)$ during the transition. Otherwise, k could not catch up to its moving target, k^*.[6]

The results for the transitional behavior of k again tell us about convergence across economies. As before, convergence depends on whether different economies have the same or different steady states. Figure 5.4 shows a case in which two economies have the same steady-state paths, k^*. Economy 1 begins at capital per worker $k(0)_1$ and economy 2 at the higher capital per worker $k(0)_2$. The graph shows that k_1 and k_2 converge toward the path k^* and that k_1 also converges toward k_2. Therefore, economy 1 grows faster than economy 2 during the transition to the steady-state path. In other words, if the two economies have the same steady-state paths, absolute convergence holds, and the poorer economy (with lower $k[0]$) grows faster. These results are similar to those found in Figure 4.7 for the model without technological progress, where $g = 0$.

Figure 5.5 considers the case in which the two economies have different steady-state paths. We assume that economy 1—with lower $k(0)$—also has a lower steady-state path, k^*. We discussed in chapter 4 why an economy with low k^* would tend also to have low k when observed at an arbitrary time, such as date 0. The graph shows that each economy converges over time toward its own steady-state path—k_1 toward k_1^* and k_2

---

[5] More precisely, the steady-state path will be a straight line, as shown in the figure, if we graph the logarithm of k^* versus time. The appendix shows how to determine k^* algebraically.
[6] We are assuming that $k(0)$ lies below the steady-state path, k^*, as shown in Figure 5.2.

toward k_2^*. Since $k_1(0)$ is less than $k_2(0)$ and k_1^* is less than k_2^*, we cannot be sure which economy grows faster during the transition. The lower $k(0)$ tends to make economy 1 grow relatively fast, but the lower k^* tends to make it grow relatively slow. Thus, convergence need not hold in an absolute sense. However, conditional convergence still applies—if we hold fixed the steady-state path, k^*, a lower $k(0)$ leads to higher growth during the transition.

C. Endogenous technological progress

The inclusion of exogenous technological progress allows the Solow model to match the observed long-term growth rates of real GDP per person observed in the United States and other advanced countries. However, this fix of the model has been criticized by many economists because the technological progress comes from nowhere—it is not explained within the model. For that reason, economists led by Paul Romer in the late 1980s and early 1990s tried to extend the model to explain why technological progress occurred. The models that Romer and others developed are called **endogenous growth models** because, first, the model explains the rate of technological progress, and, second, as in the Solow model, the technological progress leads to long-run growth of real GDP and capital per worker.

Most endogenous growth models focus on business's investments in **research and development** or **R&D**. In these models, successful R&D projects lead to the discovery of new products, better products, or superior methods of production. In the Solow growth model, we can think of these research successes as increases in the technology level, A. However, in contrast to the Solow model with exogenous technological progress, the growth rate of *A* is explained within the model. This endogeneity of technological progress means that we can use the model to understand how government policies and other variables influence R&D investment and, thereby, the rate of technological progress and the long-run growth rate of real GDP per person.

Theories of technological progress specify a connection between R&D investment and the amount of technological advance, represented by increases in *A*.[7] Since research entails discovery, the outcomes are uncertain. For example, when working on new medicines, computer designs, or other original products or processes, a researcher would not know the degree of success in advance. The uncertainty would be greater for basic research than for refinements of existing products or methods of production. However, we can say generally that more R&D investment leads to a larger expected increase in the technology level, *A*. Therefore, to have more technological progress on average, the individuals or businesses that do the innovating must be motivated to raise their R&D outlays.

In many respects, R&D investment resembles the familiar investment in physical capital. The R&D outlays correspond to investment expenditures, and the technology level, *A*, corresponds to the stock of capital, *K*. However, there are two important differences between technological progress and increases in the stock of capital. One has to do with **diminishing returns** and the other with **ownership rights**.

A key question is whether diminishing returns apply to R&D investment. Specifically, as the technology level, *A*, grows, does it become more costly in terms of

---

[7] A different approach, pioneered by Ken Arrow (1962), assumes that producers become more efficient with existing techniques through experience or, as it is sometimes called, learning-by-doing. This learning by doing has been shown to be important in the production of ships and airplanes but has not been applied empirically to technological progress for a whole economy.

R&D outlays to generate further increases in A? If so, the R&D process exhibits diminishing returns, and it may be impossible for R&D investments to sustain technological progress and long-run growth of real GDP per worker. If not, it may be possible for R&D investments to maintain technological progress and long-run growth of real GDP per worker.

To understand the second distinction between technology and the stock of capital, think of the technology level, A, as representing an idea about how to use factor inputs, K and L, to produce output, Y. In contrast, think of the stock of capital, K, as a machine or a building. If A represents an idea, it can be used simultaneously by all producers in an economy. If producer 1 uses the idea to create goods and services, producer 2 can use the same idea at the same time to create other goods and services. In a physical sense, an idea is **non-rival**—it can be used simultaneously by any number of producers without reducing the amount of the idea available to others. Examples of non-rival ideas are mathematical formulas in calculus, chemical formulas for drugs, codes for computer software, and the notes in a song. The important point about a non-rival idea is that, once it has been learned, it would be efficient to share the idea with all potential users.

The stock of capital is different from the stock of ideas. If one business uses a machine to produce goods, it is physically impossible for other businesses to use the machine at the same time. The same property holds for labor input and most other goods and services. Economists describe these goods and services as **rival**.

Suppose, however, that all ideas were freely available once discovered. In this case, few individuals or businesses would devote resources to making inventions. The learning of the ideas typically requires R&D investment, but there would be no individual payoff for an innovation. As an example, the discovery of a chemical formula for a new drug typically entails substantial R&D outlays. If the formulas for the successful drugs were distributed freely and if all firms were allowed to use these formulas without charge, there would be no way for the innovating company to recoup its research expenses. Then—at least if we are relying on profit-seeking private enterprises—little R&D would take place and little technological progress would occur.

Profit-seeking companies undertake R&D investment only if these companies can maintain some form of rights in the (good) ideas that they discover. These rights are called **intellectual property rights**. In some areas, intellectual property rights can be maintained through **patents** (typically for 17 or 20 years) and **copyrights** (usually for the lifetime of the author plus 50 years). These legal protections are especially important for pharmaceuticals, software, books, music, and movies. Some of the issues in these areas are discussed in the nearby case study of Napster and Prozac.

The story of Napster and Viagra

What do Napster (the once popular Internet site for copying music) and proposals to limit prescription drug prices have in common? Both seek to reduce prices of goods that cost little to produce now but were expensive to create initially. Cutting prices today looks great for users and, arguably, for society as a whole. If it costs virtually nothing to copy a CD over the Internet, why should people not be able to copy and listen to the music, rather than having to pay $15 at the local store? If it costs only a few dollars to produce and distribute a standard quantity of Viagra, why should people not be able to use the drug if they are willing to pay $10, rather than $100?

The problem is that the "high" prices are the rewards for the costly efforts that came before. Music companies and artists expend time and money to create hits, and the bulk of the expenses are for failed projects. To compensate for these efforts and to provide incentives for future hits, the industry has to reap large profits on its few successes.

Piracy is a problem for producers of music and similar products, such as books, movies, and computer software. The incentive to abridge intellectual property rights reflects the big gap between the prices charged by the copyright owners and the actual costs of copying and distribution. Innovations in the Internet and computer technology have dramatically lowered these costs. On the one hand, these advances are desirable, because they allow products to reach a vastly expanded audience. On the other hand, the down side is the threat to intellectual property rights. These rights are partly a matter of fairness, in the philosophical sense that inventors ought to be able to control the use of their discoveries. But, more concretely, if intellectual property rights disappear and no other effective method of compensating creativity is adopted, we will see much less of future greatness in music, books, movies, and software.

It may be that the Internet makes impossible the effective enforcement of intellectual property rights in certain areas. If so, we are likely to be in trouble with respect to future creativity. However, the best policy would be to try to maintain some degree of property rights, and the pursuit of the legal case in 2001 against Napster's Internet-based copying facility was probably a helpful part of this policy.

Prescription drugs are similar in many respects. One way to see that retail prices of patented drugs exceed current costs of production is to compare U.S. prices with the lower ones that prevail in some other countries. For example, many drugs sell in Canada for about one-half of the U.S. price. Some people conclude that the United States ought to adopt Canada's policies for pricing of prescription drugs or, alternatively, allow re-importation of the cheaper goods back to the United States. A more reasonable view is that the incentives for drug research and innovation created by high U.S. prices give Canada, Mexico, and other small markets what economists call a **free ride**. The idea of a free ride is that it allows some people—in this case, Canadians and Mexicans—to enjoy the fruits of someone's labor without having to share in the costs. Specifically, these small countries can benefit from low prices of prescription drugs without having to worry about the effects on the overall market and, hence, on the incentives for companies to develop new drugs.

The United States does not have the option to free ride because the United States is such a large part of the market for prescription drugs. If the United States were to follow Canada's lead, fewer new drugs would be available. Thus, for the United States, the choice is whether to have many effective new drugs at high prices or to have few new drugs at low prices. This choice is the relevant one for society as a whole—and, specifically, for the United States—but many people fantasize that they can have low prices AND many new drugs. Unfortunately, it just ain't so.

For many basic discoveries, patent protection is unavailable, partly because of legal limitations and partly because of practical considerations in defining the scope of an idea. For example, Isaac Newton did not have patent protection for his mathematical innovations in calculus, Solow did not have property rights in his growth model, and Henry Ford did not have exclusive use over the assembly line. For a more recent example, Toyota did not have property rights over the idea of just-in-time inventory management, and this idea was copied by other automobile manufacturers, Dell Computer, and many other companies.

In many cases, businesses that make patentable inventions do not seek patents, sometimes because the approval process is costly and, more often, because businesses do not want to reveal the information needed to gain approval. Such information tends to aid competitors even when patents are granted. In the absence of formal protection from patents, the main methods of maintaining intellectual property rights are secrecy and the advantages from moving first into a new area.

Paul Romer (1990) constructed the first model in which R&D investment and intellectual property rights were linked to a theory of technological progress and economic growth. In his model, an inventor retained perpetual monopoly rights over his or her invention. However, this extreme form of intellectual property rights was not necessary for Romer's results. The more general idea is that some form of intellectual property rights ensured that successful innovators were rewarded for their discoveries.

The Romer model distinguishes the return to society from an invention from the private return, which is the reward to the inventor. The private return is greater than zero because of the intellectual property rights, but the social return still tends to exceed the private one. For example, the social benefits from the invention of the transistor or the micro-chip were much greater than the payoffs to the individuals and businesses that made the discoveries.[8] For this reason, the resources devoted to R&D and the resulting rate of technological progress tend to be too low from a social perspective. This reasoning is often used to justify government subsidy of innovative activity, especially basic research. However, government subsidies also create problems, including the politics of choosing what to subsidize and the necessity to raise revenues to pay for the subsidies.

Romer equated technology with ideas, and he assumed that the returns from generating new ideas did not diminish as the technology level advanced. His reasoning

---

[8] Theoretically, the private returns may exceed the social ones. This situation can apply if resources are wasted when competing researchers strive to be the first to make a discovery, or if the main consequence from an improved product is the transfer of monopoly profits from the old industry leader to the new one. However, it is hard to present convincing empirical examples of this theoretical possibility.

was that the number of potentially good ideas was unlimited, so that the stock of remaining ideas would not be depleted as more things were discovered. Thus, at least as a working hypothesis, we might assume that the returns to the creation of ideas are constant. This assumption turns out to be consistent with a constant, steady-state growth rate of real GDP per worker, driven by technological progress at a constant rate. That is, the results look like those in the Solow model when the technology level, A, grows exogenously at the constant rate g.

In the Romer model, the rate of technological progress depends on the private rewards from making discoveries. These rewards depend on a number of factors:

- The private return to R&D investment is higher if the costs of R&D are lower in relation to the effects of successful innovations on sales revenues and production costs. One consideration is the size of the market over which the benefits from a discovery can be spread. A bigger market, which includes domestic and foreign sales, encourages more R&D investment.
- The private return is higher if intellectual property rights over the use of an invention are more secure and long-lasting. In many cases, these rights will be better protected domestically than internationally. Another consideration is the ease with which competitors, domestic or foreign, can imitate successful innovations. The easier the imitation the lower the intellectual property rights over an innovation and, hence, the smaller the incentive for R&D investment.
- The private return is higher if the government subsidizes R&D investment. On the other side, innovation is deterred by high costs of gaining government approval or satisfying government regulations.

Changes in any of these factors influence the rate of technological progress and, therefore, the economy's steady-state growth rate of real GDP per worker. These effects are analogous to those from changes in the rate of exogenous technological progress, *g*, in the Solow model.

R&D expenditures turn out empirically to be concentrated in the most advanced countries. This concentration also applies to numbers of patents granted and to numbers of scientists and engineers working in a country. These patterns suggest that basic innovations result mostly from innovative activities in the rich countries.[9] The concentration of R&D investments in the advanced countries makes sense because these investments are worthwhile mainly in these countries. One reason is that the rich countries have complementary resources to support research, including a supply of highly skilled workers and strong educational institutions. Another consideration is that a large market is critical for justifying the costs of innovating. Hence, the large domestic market available to wealthy countries is an important consideration. However, a small country can be successful at innovative activities if the country is well connected to other markets through international trade and if intellectual property rights are respected in foreign countries. As examples, Sweden and Finland have been leaders in pharmaceuticals and telecommunications.

At the start of this chapter, we discussed cross-country empirical research on convergence and other aspects of economic growth. These empirical findings match up well with the version of the Solow model that includes exogenous technological progress. Thus far, less cross-country empirical work has been done on the endogenous growth

[9] India is an exception as a poor country that contributes substantially to developments in computer software.

models. One finding, however, is that countries that expend more on R&D investment tend to have higher growth rates of real GDP per worker.[10]

D. The diffusion of technology

For the world as a whole, the only way to raise the technology level, *A*, is for someone to discover something new. However, for an individual country or producer, it is possible to raise *A*—the technology level available to that country or producer—by imitating or adapting someone else's innovation. For example, color television sets were invented in the United States by RCA (Radio Corporation of America), but the technology for producing televisions was copied (and improved upon) in Japan. Similarly, the technology for operating steel mini-mills in the United States was based on innovations in Germany and elsewhere.

We use the term **diffusion of technology** to describe the imitation and adaptation of one country's technology by another country. For low-income countries, this imitation and adaptation tends to be less expensive than invention as a way to improve methods of production or to introduce new and better products. Therefore, low-income countries tend to focus on diffusion of technology as the way to raise technology levels.

Several methods have been used to imitate leading technologies. In one case, a multi-national firm from an advanced country uses an advanced technology in a foreign subsidiary. Domestic entrepreneurs can then learn from the foreign-owned operations about products and production processes. These channels of technological diffusion were

---

[10] See Coe and Helpman (1995). For evidence on the relation between R&D and productivity at the level of industries and firms, see Griliches (1998).

important in the textile industries in Hong Kong and Mauritius (an economically successful island off the east coast of Africa).

Sometimes the transfer of technology occurs through observation and analysis of products exchanged in international trade. For example, an importer of a good may be able to deduce how the good was produced by taking it apart (through a process of "reverse engineering"). In other cases, a foreign company licenses or sells its processes to domestically owned businesses. For example, Nucor—the first steel mini-mill producer in the United States—purchased technological designs from a German company. In still other cases, domestic residents work or study at a business or university in an advanced country and bring back information about technology to their home countries.

The diffusion of technology provides another reason why poor countries tend to converge toward rich ones. Low-income countries are poor partly because they have access to inferior technologies. Therefore, these countries can grow rapidly by imitating better technologies from advanced countries. However, as imitation proceeds, the supply of uncopied technologies decreases, and the cost of further imitation tends to rise. This rising cost is similar to the decreasing average product of capital, y/k, in the Solow model. Therefore, the growth rates of follower countries tend to decline, and their levels of real GDP per worker tend to converge toward those in the advanced countries.

Studies show that the rate of technological diffusion to a developing country is higher when the country has a lot of international trade with rich countries, high education levels, and well-functioning legal and political systems.[11] These factors were

[11] See, for example, Jaumotte (2000) and Caselli and Coleman (2001).

strong in some East Asian countries and, therefore, help to explain the high rates of economic growth in East Asia since the 1960s.

Hybrid corn: a case of technological diffusion

One of the first studies of the diffusion of technology was Zvi Griliches's (1957) research on hybrid corn. The basic idea about hybrid corn—crossing specially selected strains of corn to develop types suitable to local conditions—was familiar to agricultural scientists since the early 20th century. However, the first successful commercial application did not occur until the 1930s in Iowa. Additional research was then required to develop hybrids that worked well in other states. Griliches showed that the time of application to a new state depended on the costs of the necessary refinements and on the potential gains in terms of crop yields and market sizes. The speed of acceptance within a state also depended on the economic benefits from the use of the hybrids. Figure 5.6 shows when hybrid corn was introduced in various states and how fast the adoption took place within each state. One reason for the delay in the use of hybrids in the southern states was that these hybrids had to be substantially modified from those used in Iowa. Applications of models of technological diffusion to innovations in other industries include Gort and Klepper (1982) and Jovanovic and Lach (1997).

III. Summarizing What We Know about Economic Growth

We began our study of economic growth in chapter 3 with the Solow growth model. We can think of the model as having two phases. In the first phase, capital and real GDP per worker rise from their initial levels to their steady-state levels. The second phase is the steady state. In chapters 3 and 4, capital and real GDP per worker did not grow in the long run, that is, in the steady state. However, in chapter 5, the inclusion of technological progress led to growth of capital and real GDP per worker in the steady state.

In chapter 4, we used the Solow model to predict the short- and long-run effects from changes in the saving rate, the technology level, the size of the labor force, and population growth. Then we showed how the transition phase of the model predicted convergence, by which we mean a tendency of poor economies to catch up to rich economies. We found that absolute convergence conflicted with observations for a broad group of countries but that a modified concept—conditional convergence—was consistent with these observations. Conditional convergence allows for differences in steady-state positions, due to variations in saving rates, technology levels, and population growth rates. In extended models, the differences can reflect additional variables, including legal and political systems, openness to international trade, and the efficiency of education and health programs.

In chapter 5, we showed that the concept of conditional convergence helps us to understand many historical patterns of economic growth. We can explain why some war-ravaged countries grew rapidly after World War II. We can also explain why most East Asian countries grew fast from 1960 to 2000, while most sub-Saharan African countries grew slowly or not at all. Thus, our study of transitions and convergence helps us to understand growth patterns in all parts of the world and for periods as long as several decades.

The basic Solow model does not explain long-run growth of real GDP per person, as observed for well over a century in the United States and other advanced countries. We showed that the model would explain a long-term growth rate of around 2% per year if we assumed exogenous technological progress at a rate of about 1% per year. Then we considered attempts to make technological progress endogenous. These endogenous growth models rely on R&D investment as the source of improvements in technology. The models predict how intellectual property rights, research subsidies, and other variables affect the rate of technological progress and, hence, the long-run growth rate of real GDP per person.

We also considered technological diffusion, which is the main method by which low-income countries raise their technology levels. This diffusion of technology helps to explain convergence of poor countries toward rich countries but does not explain technological progress for the whole world.

Although we understand a lot about economic growth, there is much that remains unexplained. For example, economists have isolated only some of the variables that lead to differences across countries in steady-state positions. In the long-run context, we are still uncertain about the sources of technological progress. In particular, we cannot say with any assurance how government policies that affect the incentives for R&D investment would influence long-run economic growth in a single country or in the world. Thus, although we have learned a lot, there is still much to do.

Questions and Problems

Mainly for review

5.1. Suppose that the technology level, A, grows exogenously at a positive rate, $g > 0$. Does the level of output, Y, grow in the long run? Does output per worker, Y/L, grow in the long run?

Problems for discussion

5.x. Convergence and the dispersion of income (difficult)

Consider a group of economies that satisfies absolute convergence, that is, poor economies tend to grow faster than rich ones.

a. Does this convergence property imply that a measure of the dispersion of income per person—or income inequality—across the economies will narrow over time? [This question relates to Galton's fallacy, an idea applied by Galton to the distribution of heights and other characteristics in a population. If a parent is taller than average, the child tends to be taller than average but shorter than the parent. That is, there is reversion to the mean, an effect that parallels the idea of absolute convergence. Does the presence of reversion to the mean imply that the distribution of heights across the population will narrow over time? The answer is no, but you are supposed to explain why.]

b. We found in Figure 4.10 that absolute convergence held for the U.S. states from 1880 to 2000. A measure of the dispersion of per capita income across the states declined for most of the period from 1880 to 1970 (except for the 1920s and 1930s). Dispersion did not change a great deal since 1970. Can you relate these observations to your answer from part **a.**?

c. We found in Figure 4.8 that absolute convergence did not hold for a broad group of countries from 1960 to 2000. We did find in Figure 5.1 that conditional convergence held for these countries. A measure of the dispersion of per capita real GDP across these countries shows a mild, but systematic, increase from 1960 to 2000. How would you account for this pattern?

Appendix: The steady-state path in the Solow model with exogenous technological progress

We consider here the determination of the steady-state path, k^*, in the model with exogenous technological progress. The path is shown graphically in Figure 5.3. This appendix provides an algebraic derivation of k^*.

The growth rate of capital per worker is given by

(4.8) $$\Delta k/k = s\cdot(y/k) - (\delta+n).$$

Hence, along a steady-state path, the growth rate is

(5.12) $$(\Delta k/k)^* = s\cdot(y/k)^* - (\delta+n),$$

where $(y/k)^*$ is the average product of capital in a position of steady-state growth. We also know that, in a situation of steady-state growth, k grows at the rate

(5.10) $$(\Delta k/k)^* = g/(1\text{-}\alpha).$$

Therefore, if we substitute $g/(1\text{-}\alpha)$ for $(\Delta k/k)^*$ on the left-hand side of equation (5.12), we get

$$g/(1\text{-}\alpha) = s\cdot(y/k)^* - (\delta+n).$$

We can rearrange the terms to get

$$s\cdot(y/k)^* = (\delta+n) + g/(1\text{-}\alpha).$$

Then, if we divide through by s, we get a formula for the steady-state average product of capital:

(5.13) $$(y/k)^* = (1/s)\cdot[(\delta+n) + g/(1\text{-}\alpha)].$$

Note that the right-hand side of the equation is constant. Therefore, this result verifies that the average product of capital, $(y/k)^*$, is constant in a position of steady-state growth.

Recall that the production function is

$$y = A \cdot f(k),$$

so that we can write the average product of capital, y/k, as

$$y/k = A \cdot f(k)/k.$$

Therefore, if we define k^* to be the path that k follows during steady-state growth, the steady-state average product of capital is

$$(y/k)^* = A \cdot f(k^*)/k^*. \tag{5.14}$$

Equations (5.13) and (5.14) give us two expressions for $(y/k)^*$. Therefore, the two right-hand sides must be equal:

$$A \cdot f(k^*)/k^* = (1/s) \cdot [(\delta+n) + g/(1-\alpha)]. \tag{5.15}$$

The right-hand side is constant, and the technology level, A, on the left-hand side grows over time at the rate g. Therefore, if we specify the form of the production function, f, we can use equation (5.15) to determine the path k^*.

Suppose that the production function, $f(k)$, takes the Cobb-Douglas form,

$$y = Ak^{\alpha}, \tag{3.18}$$

which we discussed in the appendix to chapter 3. In this case,

$$\begin{aligned} A \cdot f(k)/k &= Ak^{\alpha}/k \\ &= Ak^{\alpha}k^{-1} \\ &= Ak^{\alpha-1} \\ A \cdot f(k)/k &= Ak^{-(1-\alpha)}. \end{aligned}$$

Therefore, we can substitute $A \cdot f(k^*)/k^* = A \cdot (k^*)^{-(1-\alpha)}$ in equation (5.15) to get

$$A \cdot (k^*)^{-(1-\alpha)} = (1/s) \cdot [(\delta+n) + g/(1-\alpha)].$$

If we multiply through by $(k^*)^{1-\alpha}$ and s, divide through by $[(\delta+n) + g/(1-\alpha)]$, and rearrange the terms, we get

$$(k^*)^{1-\alpha} = \frac{sA}{(\delta + n) + g/(1-\alpha)}. \tag{5.16}$$

On the right-hand side, everything except *A* is constant over time. If *A* were also constant, *k** would be constant, as in the Solow model without technological progress ($g = 0$). If *A* grows at the rate *g*, the equation determines how *k** varies over time.

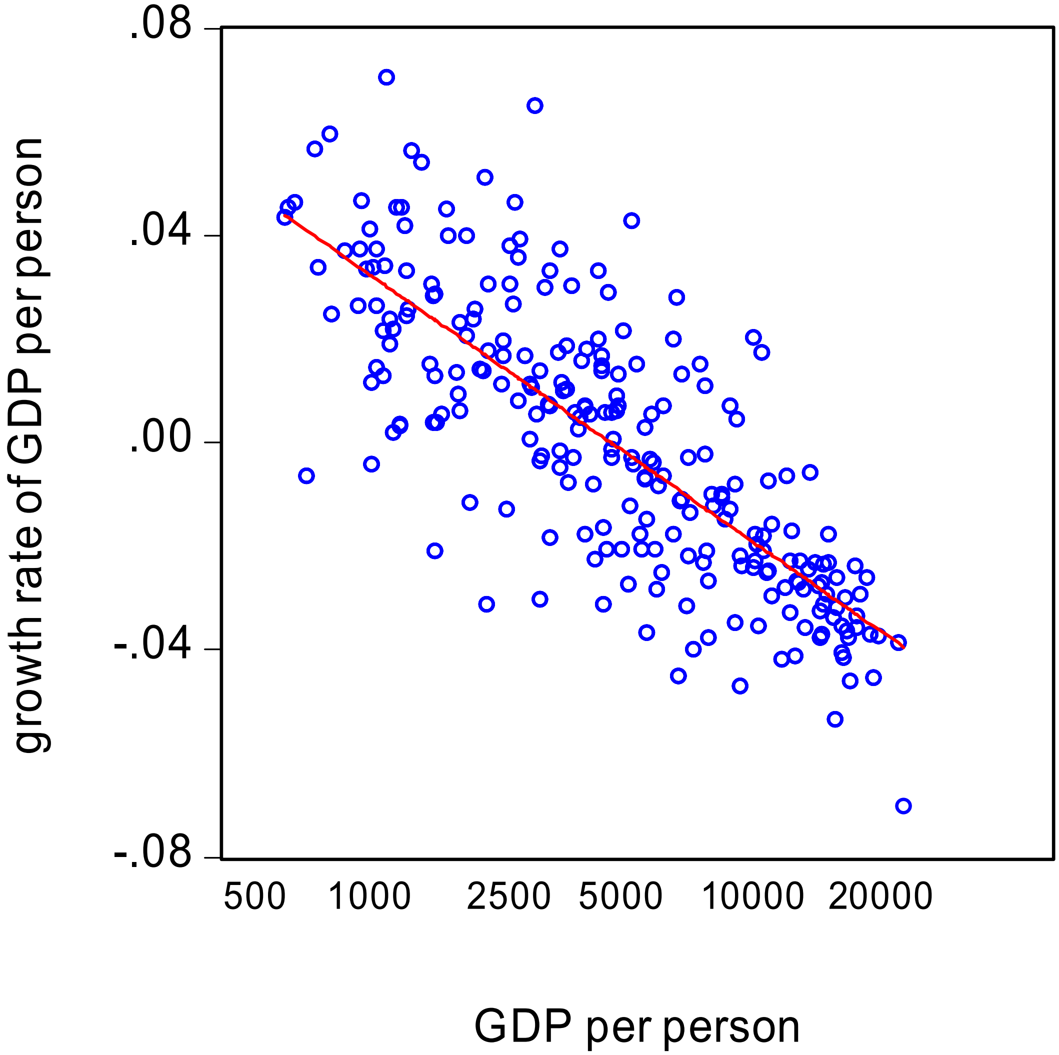


Figure 5.1
Growth Rate versus Level of Real GDP per Person:
Conditional Convergence for a Broad Group of Countries

The horizontal axis shows real GDP per person in 1995 U.S. dollars on a logarithmic scale for 76 countries in 1965, 86 countries in 1975, and 84 countries in 1985. (This sample was based on the availability of data—87 countries appear at least once.) The vertical axis shows the corresponding growth rates of real GDP per person—for 1965-75, 1975-85, and 1985-95. Each of the growth rates filters out (and, therefore, holds constant) the estimated effects of the variables discussed in the text. The red line is the straight line that provides a best fit to the relation between the growth rate of real GDP per person (the variable on the vertical axis) and the level of real GDP per person (on the horizontal axis). The line has a clear negative slope. Therefore, once we hold constant the other variables, a lower level of real GDP per person matches up with a higher growth rate of real GDP per person. This relation is called conditional convergence.

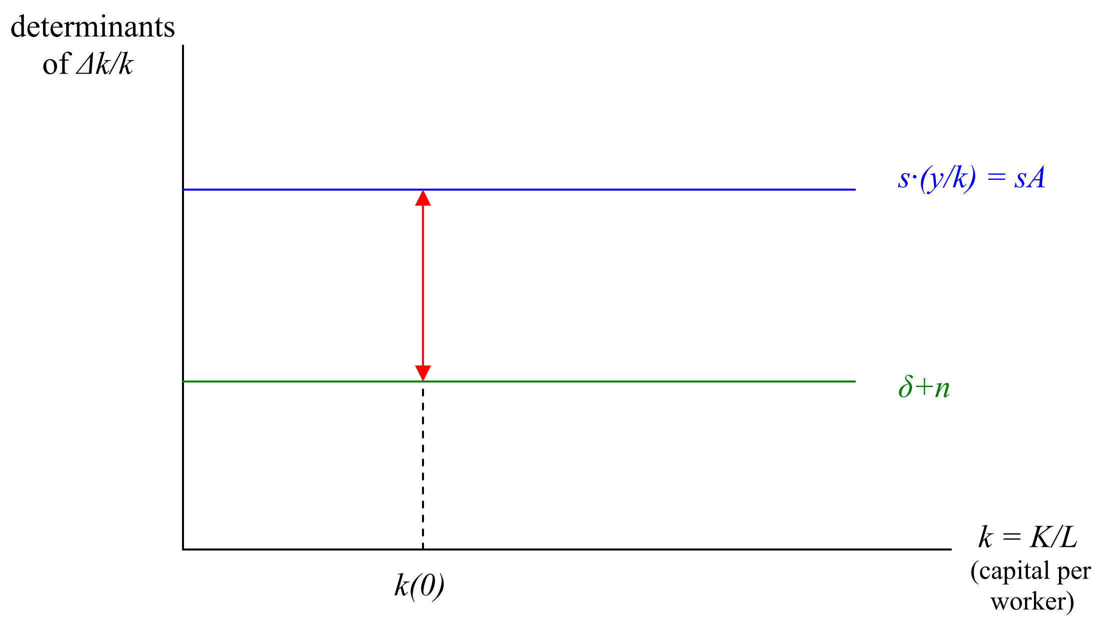


Figure 5.2

Economic Growth with Constant Average Product of Capital

This graph modifies Figure 4.4 to allow for a constant average product of capital, *y/k*. In this *Ak* model, *y/k* equals the constant technology level, *A*. Therefore, the *s·(y/k)* curve becomes the horizontal line *sA*. If *sA* is greater than $\delta+n$, as shown, the growth rate $\Delta k/k$ is a constant equal to the vertical distance between the two horizontal lines. This distance is shown by the red arrows.

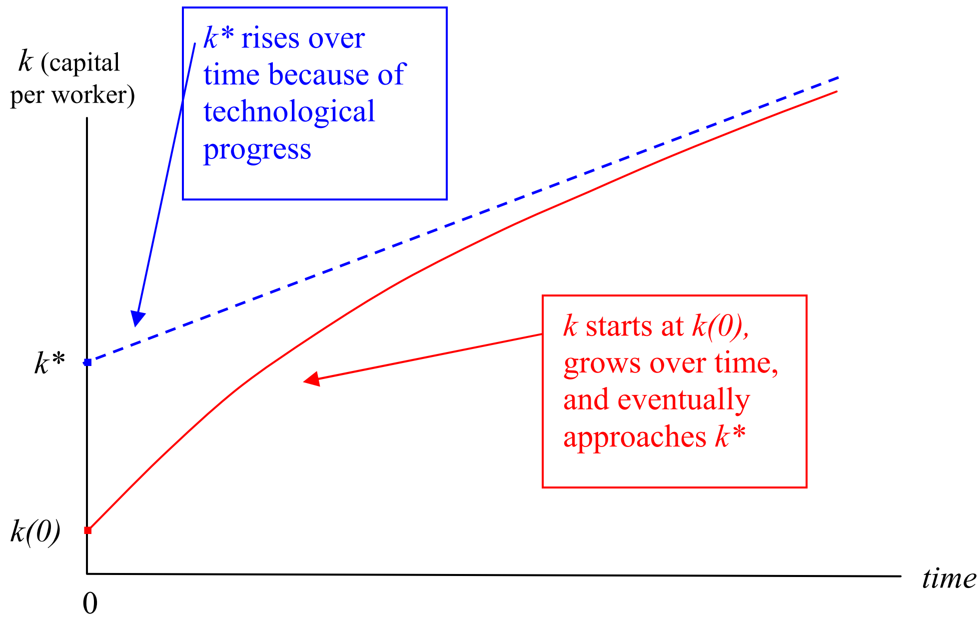


Figure 5.3

The Transition Path for Capital per Worker in the Solow Model with Technological Progress

In the Solow model with technological progress at rate g, the steady-state level of capital per worker, k^*, is not constant—k^* rises over time along the blue line. In the transition, capital per worker, k, starts at $k(0)$, rises over time along the red curve, and gradually approaches the k^* line.

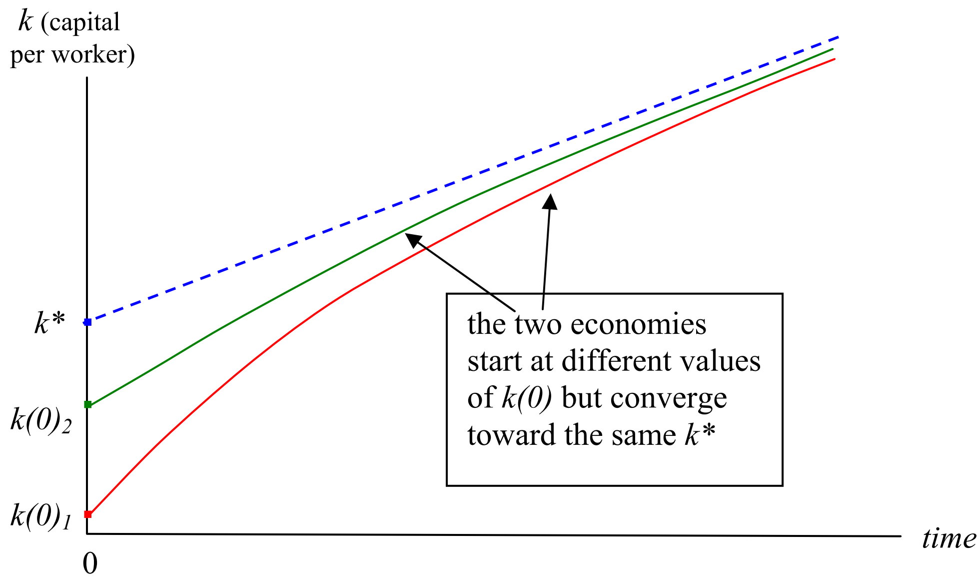


Figure 5.4

**Convergence and Transition Paths for Two Economies
in the Solow Model with Technological Progress**

As in Figure 5.3, the steady-state level of capital per worker, k^*, rises over time along the blue line. The first economy starts at $k(0)_1$, and the second economy starts at the higher value $k(0)_2$. During the transitions, capital per worker in each economy, k_1 and k_2, gradually approach the common steady-state path, k^*, shown by the dashed blue line. The first economy (the red curve) grows faster than the second economy (the green curve), so that k_1 converges toward k_2. Therefore, absolute convergence applies.

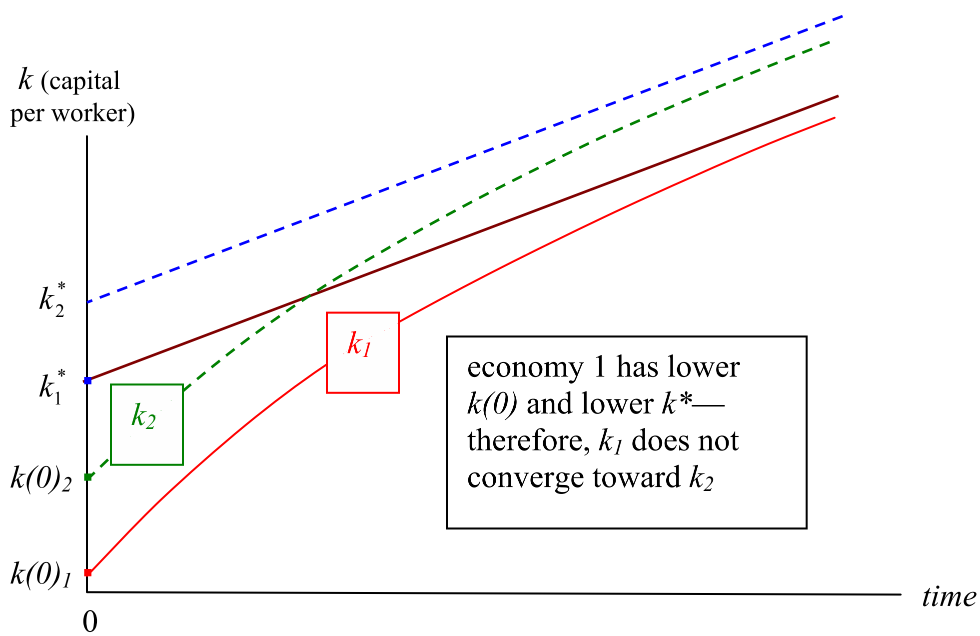


Figure 5.5

Failure of Convergence and Transition Paths for Two Economies in the Solow Model with Technological Progress

As in Figure 5.4, the first economy starts at $k(0)_1$, and the second economy starts at the higher value $k(0)_2$. However, economy 1 now also has a lower level for the path of capital per worker in the steady state, that is, the brown line for k_1^* lies below the dashed blue line for k_2^*. During the transitions, k_1 and k_2 gradually approach their respective steady-state paths, k_1^* and k_2^*. However, economy 1 need not grow faster than economy 2, and k_1 (the red curve) does not converge toward k_2 (the dashed green curve). Hence, absolute convergence need not hold.

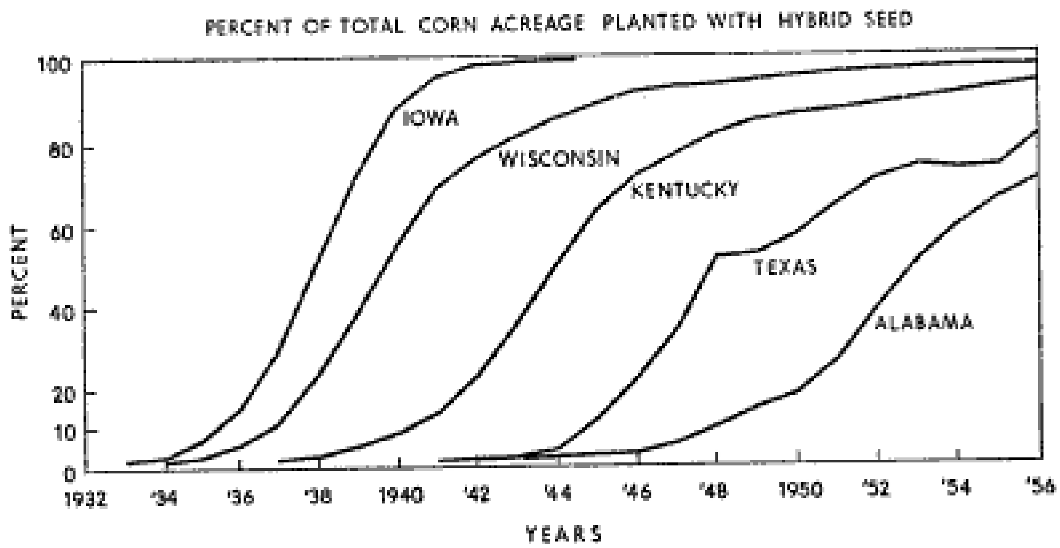


Figure 5.6

Technological Diffusion for Hybrid Corn

The innovations in hybrid corn began in the United States in Iowa in the early 1930s. This innovation spread later to other agricultural states, as shown in the graphs. Each curve shows the fraction of corn acreage planted in various years with hybrid seed.

January 12, 2005

Part III: Economic Fluctuations

People care a lot about whether the economy is expanding or contracting. During an **economic boom**, when real GDP rises, consumption and investment tend to be strong, employment tends to rise, and unemployment tends to fall. Conversely, during a **recession**, when real GDP falls, consumption, investment, and employment tend to be weak, and unemployment tends to increase. One indicator that people are happier during booms than during recessions is that the first President George Bush lost the U.S. presidential election in 1992 in the aftermath of the recession of 1990-91. Similarly, Richard Nixon attributed his loss of the presidency to John Kennedy in 1960 to the weak economy (while Nixon was vice president during the Eisenhower administration). Conversely, Bill Clinton was probably reelected as president in 1996 because of the strong economy.

In this part of the book, our goal is to understand these **economic fluctuations**, that is, the increases of real GDP during booms and the decreases during recessions. As part of this study, we want to know how important variables other than GDP—consumption, investment, employment, wage rates, interest rates, and so on—behave when real GDP is rising during booms or falling during recessions. These economic fluctuations are usually important for relatively short periods, such as one or two years. In contrast, our study of economic growth in chapters 3-5 focused on the long term—5-10 years or even 20-30 years or longer.

To illustrate some general features of our models of economic fluctuations, we can start with the building block from chapter 3 that was central to our model of economic growth, the production function:

(3.1) $$Y = A \cdot F(K,L).$$

Thus, real GDP, Y, depends on the technology level, A, and the quantities of capital and labor, K and L. When we studied economic growth, we found from growth accounting that the growth rate of Y reflected growth in its three determinants—A, K, and L. We can see, by analogy, that short-term increases or decreases in Y have to reflect short-term increases or decreases in its three determinants—*A, K, and L*.

We start in chapters 6 and 7 with the unrealistic assumption that the quantities of capital and labor, K and L, are fixed in the short run. In that case, fluctuations in real GDP, Y, result solely from shifts to the technology level, A. Equation (3.1) tells us that changes in Y match up mechanically with changes in A if K and L are not changing. However, we will have to rely on our economic model to work out the changes in other macroeconomic variables, such as consumption and investment. We shall also use the model to derive the changes in the real wage rate, the real rental price of capital, and the interest rate.

In Chapter 8, we make the model more realistic and relevant by allowing for variations in the supplies of labor and capital. With a fixed stock of capital, variations in the supply of capital services come from changes in the utilization rate for capital. In this setting, fluctuations in real GDP are no longer tied mechanically to the changes in the technology level, A. We can also assess how changes in labor and capital input match up with the changes in real GDP during economic fluctuations.

Chapter 6 lays out the basic structure of the model. Chapter 7 uses the model to understand economic fluctuations when labor and capital are in fixed supply. Chapter 8 extends the model to allow for variations in the supplies of labor and capital services.

Chapter 6

Microeconomic Foundations and Market Clearing

To build a model of economic fluctuations, we start by working out the **microeconomic foundations**. By microeconomic foundations, we mean an analysis of how individual consumers and producers make choices. Although macroeconomic models differ on the specifics of their microeconomic foundations, every reasonable macroeconomic model must have some kind of microeconomic foundation. This chapter begins the construction of the microeconomic foundations for our model.

An example of a microeconomic choice is the decision that each worker makes about how much to work. Another choice by producers is how many workers to hire. For all of these decisions, an individual worker or producer takes as given the prices that it faces. One of these prices is the real wage rate, which specifies the quantity of goods that can be bought with an hour of labor.

We should stress that a key assumption is that individual workers, consumers, and producers are too small to have a significant impact on the various prices that influence their decisions. To take a concrete example—which we will detail later—consider a simple analysis of the labor market. Suppose that, in choosing how much labor to supply, each worker takes as given the real wage rate. Similarly, in deciding how much labor to demand, each producer takes as given the real wage rate. Thus, the individual choices of quantities supplied and demanded are made at given market prices. Economists say that this assumption applies under **perfect competition**. Under perfect competition, each market participant assumes that he or she can sell or buy any quantity desired at the going price. In particular, each participant is small enough that changes in his or her supply and demand have a negligible impact on the market price.

When we add up the individual choices, we determine aggregate or market supply and demand functions. For example, we determine the market supply of and demand for labor as a function of the real wage rate, *w/P*. Figure 6.1 shows this case. The aggregate quantity of labor supplied, L^s, is assumed to rise as w/P increases (along the blue curve). The aggregate quantity of labor demanded, L^d, is assumed to decline as *w/P* rises (along the red curve).

Once we know the market supply and demand functions, we have to consider how these functions determine the quantities and prices in the economy. Our main approach relies on **market-clearing conditions**. As an example, we have in Figure 6.1 that the market supply of and demand for labor each depend on the real wage rate, *w/P*. Our assumption is that *w/P* adjusts to clear the labor market, that is, to equate the quantity of labor supplied to the quantity demanded. Thus, the market-clearing real wage rate is the value *(w/P)** shown in the figure, and the market-clearing quantity of labor is the value *L** shown in the figure.

We now start the construction of the microeconomic foundations for our macroeconomic model. We begin by specifying the structure of the markets in the model.

I. Markets in the Macro Economy

Our macroeconomic model will have several markets on which exchanges occur. In this section, we describe the participants in each market and identify the goods and services exchanged on each market.

We simplify by assuming that households perform all of the functions in the economy. Each household runs a family business and uses labor, L, and capital, K, to produce goods, Y, through the production function, which we have seen before:

$$Y = A \cdot F(K,L). \tag{3.1}$$

More realistically, the production of goods might take place in a large corporation or a small business. However, if we included these private businesses in our model, we would have to take into account that they must ultimately be owned by households. When we think of businesses as parts of households, we avoid a lot of complexity involved with ownership structure. Since we end up with the same macroeconomic results, this simplification is worth making.

In the real world, the typical household uses little of the goods that it helps to produce in the marketplace. Usually, a person works on one or a few products and receives income from the sale of these products or from the sale of labor services, which help to produce the products. The person then spends this income on an array of goods. The model would become too complex if we tried to capture this variety of goods. Therefore, we simplify by imagining that households first sell all of the goods that they produce on a **goods market**. Then households buy back from this market the goods that they want. One reason that a household buys goods is for consumption. Another reason is to increase the stock of goods in the form of capital—machines and buildings—used for production. This use of goods is called investment.

Households supply labor on a **labor market**. To simplify at the outset, we assume that the quantity supplied, L^s, is a constant, L. This assumption is not harmless, and we shall eventually have to change it, but we wait for that until chapter 8. In terms of Figure 6.1, our assumption amounts to neglecting the upward slope of the labor-supply curve. As in previous chapters, we measure labor as the flow quantity of person-hours per year. For example, if a person works 40 hours per week, 52 weeks per year, the flow of person-hours per year is 2080.

Households, as managers of family businesses, also demand labor in the quantity L^d from the labor market. The labor demanded is used as an input to the production of goods. Notice that each household wears two hats in our simplified economy. When wearing the first hat, the household supplies labor and, thereby, looks like an employee, hired by the person who buys the labor. When wearing the second hat, the household demands labor and, thereby, looks like an employer, who hires the person who sells the labor.

Next we consider the capital input to production. Households own the capital stock, K. An individual household can add to its stock by buying goods from the goods market and can lower its stock by selling goods on the goods market. We can think of these sales as resales of used capital goods. For example, a household might sell a used automobile or house, which are forms of capital goods. In our model, where households run businesses, we also imagine that households might sell a used machine or a whole used factory.

The stock of capital, K, is a stock of capital measured in units of goods—for example, numbers of automobiles or numbers of machines. Conceptually, we should distinguish these stocks of goods from the flow of capital services. For example, suppose that a household owns a single machine. If the machine operates 8 hours per day, 5 days per week, 52 weeks per year, the machine is used for 2080 hours per year. We think of these 2080 machine-hours per year—a flow variable—as the quantity of capital services. This flow is analogous to the flow of labor services, measured in person-hours per year.

We simplify at the outset by assuming that each unit of capital—say, each machine—is used for a fixed number of hours per year, perhaps 2080. In this case, the flow of capital services is a constant multiple of the capital stock—each machine represents 2080 machine-hours per year. Therefore, in this case, we will not get into any trouble if we enter the capital stock, K, as an input to the production function, as we did in equation (3.1). This stock really represents the flow of capital services, but this flow is a fixed multiple of the stock. Since the multiple is a constant, say 2080, we do not have to show it explicitly in the production function, $F(K,L)$.

Although a household owns a particular unit of capital—say, a machine—it does not necessarily use that capital for its own production of goods. Rather, we assume that the household can rent the capital to another household, which then uses it as an input to production. For example, a household might rent its house or automobile to another household. In our model, we extend this idea of rentals to all types of capital, including machines and factories.

We shall find it convenient to assume that each household rents out all of the capital that it owns on a **rental market**. Thus, if a household owns a machine, it offers for rent all of the capital services—say, 2080 machine-hours per year—that this machine provides. In the real world, we can think of Hertz Rental Car as owning automobiles and renting them to users. Other real-world examples are rentals of furniture from a company such as Cort Furniture, rentals of tools from a store such as Home Depot, and rentals of taxis and truck cabs from their owners. The important assumption in our model is that households do not allow any of their capital to sit idle and, rather, provide all of it for use on the rental market.

The amount of capital offered on the rental market is analogous to the amount of labor offered or supplied to the labor market, L^s. Therefore, we think of the capital offered on the rental market as the supply of capital services, K^s. Since we have assumed that each household rents out all of its capital, we have $K^s = K$. (More precisely, we should multiply by 2080 to convert from capital stock—numbers of machines—to capital services—machine-hours per year—but since 2080 is a constant, we can ignore it.)

So far, we have assumed that each household rents out all of the capital that it owns. However, households, as managers of family businesses, will also want to use capital services to produce goods. To get this capital input, households rent it from the rental market. The amount of capital rented on the rental market is analogous to the amount of labor purchased or demanded from the labor market, L^d. Therefore, we think of the capital rented on the rental market as the demand for capital services, K^d.

Notice that we are assuming that each household rents out all of its capital, $K^s = K$, and then rents back the quantity K^d. If K^s is greater than K^d, we could think instead of the household as retaining the quantity K^d of its own capital for use in production of goods and then renting out only the net amount, $K^s - K^d$. Analogously, if

K^s is less than K^d, we can think of the household as using the entire quantity K^s of its own capital for use in production of goods and then renting an additional amount, $K^d - K^s$. The results would be the same under these alternative assumptions. Therefore, we shall find it convenient to stick with the assumption that the household rents out all of its capital, $K^s = K$, and then rents back the quantity, K^d, that it decides to use as an input to production.

The last market that we introduce is one in which households borrow or lend. A borrowing household receives a loan from another household, whereas a lending household provides a loan to another household. In the real world, this lending and borrowing would typically occur through financial institutions, such as banks. However, as in our neglect of private businesses, we simplify by assuming that households carry out all lending and borrowing directly.

We assume that a household that makes a loan receives a piece of paper—a form of contract—that specifies the terms of the loan. We call this piece of paper a bond, and we call the market on which households borrow or lend the **bond market**. The holder of a bond—the lender—has a claim to the amount owed by the borrower.

II. Money as a Medium of Exchange

Households buy and sell goods on the goods market, labor on the labor market, capital services on the rental market, and bonds on the bond market. We assume that the exchanges on each of these markets use a single form of **medium of exchange** or **money**. The money in our model is analogous to a paper **currency** issued by a government.

Common currency

Historically, most governments have issued their own currency. However, there is now a trend toward the formation of groups of countries that share a **common currency.** These groups are called **currency unions**. The most important example of a currency union since 1999-2001 is the twelve Western European countries that use the same money, the euro, issued by the European Central Bank. These countries are Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, and Spain. Some other European countries are now considering the adoption of the euro, although the idea has, thus far, been rejected by the United Kingdom, Denmark, and Sweden.

A number of small countries have been joined together in currency unions for some time. Examples are the 15 African countries that use the CFA franc (linked most of the time to the French franc) and the 7 Caribbean countries in the Eastern Caribbean Currency Area (ECCA) that use the Caribbean dollar (linked now to the U.S. dollar). At the current time, proposals exist for the creation of new currency unions in northeast Asia (China, Japan, and South Korea), southern and western Africa, in the Persian Gulf, in Central America, and between Australia and New Zealand. In addition, a number of small countries use the currency of a large country—examples are Panama, Ecuador, Bermuda, Liechtenstein, Luxembourg, and San Marino.

Money is denominated in an arbitrary unit, such as a "dollar." For example, a household may have $100 of U.S. currency. We refer to dollar amounts as **nominal** magnitudes. Thus, $100 is the value of the household's currency in dollar or nominal units. One important property of paper money is that it bears no interest. That is, if a household has $100 of money and just leaves it under the mattress, the amount of money held remains at $100 next week and next year (assuming that it is not lost). In contrast, bonds will earn interest.

We use the symbol M for the dollar quantity of money that a household holds. The sum of the individual holdings of money equals the aggregate quantity of money in the economy. We assume, for now, that this aggregate quantity of money is a given constant. Therefore, the total money held by all of the households must end up equaling this constant.

III. Markets and Prices

The key macroeconomic variables in our model will be determined by the interactions of households who trade on the various markets. In the model, these markets are for goods, labor, capital services, and bonds. We now describe the details of each market.

A. The goods market

We assume that there is a single type of good, which can be used for consumption or investment. The goods market is the place in which households exchange goods for money. The price in this market, denoted by *P,* expresses the number of dollars that exchange for one unit of goods. We call *P* the **price level**. The **consumer price index** or **CPI**, which we discussed in chapter 2, is a real-world example of the price level. The

CPI measures the dollar cost of a representative market basket of goods and services.[1] Alternatively, we can think of the deflator for the gross domestic product (the GDP deflator), which is a price index related to the economy's overall production of goods and services (the real GDP). In the model, which has just one type of good, the price level, P, corresponds to the CPI or the GDP deflator. We assume, for now, that P does not change over time. That is, we ignore **inflation**.

Recall that households produce goods in the flow quantity Y per year, where Y is given from the production function as

$$Y = A \cdot F(K,L). \tag{3.1}$$

Since all of these goods are sold on the goods market, the variable Y will also represent the quantity of goods per year sold and bought on the goods market. The quantity PY is the dollar value per year of the goods bought and sold on the goods market.

For a seller of goods, the price level, P, is the number of dollars obtained for each unit of goods sold. In contrast, for a buyer, P is the number of dollars paid per unit of goods. Since P dollars buy 1 unit of goods, \$1 would buy $1/P$ units of goods. The expression $1/P$ is, therefore, the value of \$1 in terms of the goods that it buys. Similarly, \$$M$ exchange for

$$(M) \cdot (1/P) = M/P$$

units of goods. Whereas the quantity M is the value of money in dollars, the quantity M/P measures the value of this money in terms of the goods that it buys. Expressions like M/P are in **real terms**—that is, in units of goods—whereas a quantity like M is in dollar or nominal terms.

[1] More precisely, the CPI measures the dollar cost of a market basket of goods in a particular year, say 2004, expressed relative to the dollar cost of the market basket of goods in a base year, say 1996.

As an example, if a household has \$100 of money and the price level is 5, the real value of the household's money is

$$100/5 = 20.$$

That is, the household could buy 20 units of goods with its \$100 of money. Hence, 100 is the dollar or nominal value of the household's money, whereas 20 is the real value of this money, measured in terms of the quantity of goods that it buys. To put it another way, each \$1 of money can buy 1/5th of a unit of goods. Hence, 1/5th is the real value of each dollar.

B. The labor market

The labor market is the place in which households buy and sell labor at the dollar or **nominal wage rate**, w. Since we measure labor, L, in units of hours worked per year, the wage rate, w, has the units of dollars per hour worked. A household that buys the amount of labor L^d pays the nominal amount wL^d per year and then gets to use the labor as an input to production. A household that sells the quantity of labor L^s receives the nominal wage income of wL^s per year.

The **real wage rate** is w/P. This real wage rate is the value in goods per hour received by a supplier of labor and paid by a demander of labor. For example, if the dollar wage rate is $w = \$10$ per hour and the price level is $P = 5$, the real wage rate is

$$w/P = 10/5 = 2.$$

This real wage rate—2 goods per hour worked—tells us the quantity of goods that can be bought with the nominal wages (\$10) paid for an hour of work. Since people care about the goods that they can get, we shall find that household decisions depend on the real wage rate, w/P, rather than the nominal wage rate, w.

C. The rental market

In the rental market, households rent out capital, K, for dollars at the dollar or **nominal rental price**, R. The price R is expressed in dollars per unit of capital per year. For example, if R = $100 per year, a household receives $100 per year for each unit of capital (say, a machine or an automobile) that the household rents out on the rental market.

A household that rents the amount of capital K^d pays the nominal amount RK^d per year and then gets to use the capital as an input to production. A household that rents out the quantity of capital K^s receives the nominal rental income of RK^s per year.

The **real rental price** is R/P. This real rental price is the value in goods per unit of capital per year received by the supplier and paid by the demander. For example, if the dollar rental price is $R = \$100$ and the price level is $P = 5$, the real rental price is

$$R/P = 100/5 = 20.$$

This real rental price—20 goods per unit of capital per year—gives the quantity of goods that can be bought with the rental payments ($100) for each unit of capital over a year. Again, since people care about the goods that they can get, we shall find that household decisions depend on the real rental price, R/P, rather than the nominal rental price, R.

D. The bond market

Our model has a simple form of bond market in which households borrow and lend from each other. For example, a household might make a loan to another household that wants to buy a car or house or a machine or factory. We call the piece of paper that lays out the terms of the loan contract a **bond**.

A bond could be an I.O.U. that says that household *a* owes a certain number of dollars to the holder of the bond. Initially, the borrower owes the money to household *b*, which is the household that advanced the money. However, we assume that bonds can be sold on the bond market to another household, perhaps household *c*, which becomes the holder of the bond. Household *a* then owes the money to household *c*.

We define units so that each unit of bonds commits the borrower to repay $1 to the holder of the bond. This $1 is the **principal** of the bond.

We simplify by thinking of all bonds as having very short **maturity**. At any point in time, the issuer of a bond—the borrower—is entitled to buy back the bond for the fixed $1 of principal. That is, the borrower can retire the loan by giving back the $1 to the holder of the bond. Similarly, the holder of the bond is entitled to return the bond to the borrower at any time in exchange for $1. That is, the holder can cancel the loan by demanding the $1 of principal.

We assume that, as long as a bond is neither retired nor cancelled, each unit of bonds commits the borrower to pay the holder a flow of **interest** payments of *$i* per year. The variable *i* is the interest rate, which is the ratio of the interest payment, *$i*, to the principal, *$1*. The interest rate, *i*, can vary over time.

For the holder of a bond, the interest rate, i, determines the return per year to lending. For the issuer of a bond, i determines the cost per year of borrowing. Since we have assumed that the price level, P, is constant, i is both the **nominal interest rate** and the **real interest rate**. We distinguish these two interest rates in chapter 11 when we consider inflation.

We simplify by assuming that all bonds are alike, regardless of the household that issued the bond. Most importantly, we neglect differences among issuers in the **risk** that payments of interest and principal will not be made. Given this assumption, the interest rate, i, will have to be the same on all bonds. Otherwise, borrowers would want to do all their borrowing at the lowest interest rate, whereas lenders would want to do all their lending at the highest interest rate. Since all bonds are identical, borrowers and lenders can be matched only if the interest rates on all bonds are the same.

Let B represent the number of bonds in dollar or nominal units that a household holds. This amount may be greater than zero or less than zero for an individual household. Notice, however, that for any dollar borrowed by one household, there must be a corresponding dollar lent by another household. Hence, the total of positive bond holdings for lenders must exactly match the total of negative bond holdings for borrowers. Therefore, when we sum up over all households, the total of the B's must always be zero.

Finally, consider the price of bonds. One unit of bonds was defined to have a principal of *\$1*—that is, each unit can always be exchanged for *\$1* by canceling the loan. Therefore, the nominal price of these bonds must always be *\$1* per unit.[2] However, the

[2] The price is fixed at *\$1* per unit because we are considering bonds with very short maturity. Longer term bonds commit the borrower to pay the bond holder a stream of nominal payments (in the form of coupons

important variable for our analysis will be the interest rate, i. We can think of i as the cost or price of credit. A higher i means that obtaining credit—that is, borrowing—is more expensive in terms of the interest that has to be paid. At the same time, a higher i means that extending credit—that is, lending—is more rewarding in that it yields a higher flow of interest income.

IV. Income, Expenditure, Assets, and the Budget Constraint

The quantities and prices determined on the four markets will determine household income. Households will receive income from managing the family business, wages, rentals of capital services, and interest received. These flows of income are **sources of funds** for households. Households use their sources of funds to buy goods or to increase their assets, that is, to **save**. The purchases of goods and assets are **uses of funds** for households. The important point is that the total sources of funds must equal the total uses of funds. This equality is called the household **budget constraint**. In this section, we derive the household budget constraint. In chapter 7, we use the budget constraint to understand how households make decisions about consumption and saving.

A. Income

Begin by considering household income. Households receive income in four forms: profit from the family business, wage income, rental income, and interest income. We consider each of these in turn.

---

and principal) over a period up to the maturity date. We can define the units so that each unit of long-term bonds commits the borrower to pay the holder *$1* at the maturity date. The nominal price of these long-term bonds will vary when the interest rate, i, changes.

1. Profit. Households earn profit from their business activities. If a household uses the quantity of labor L^d and the quantity of capital K^d as inputs to production, the amount of goods produced, Y, is given by the production function:

(6.1) $$Y = A \cdot F(K^d, L^d).$$

Since all goods sell at the price level, P, the nominal income from sales is PY per year.

Households pay the nominal amounts wL^d per year for labor input and RK^d per year for capital input. The difference between the income from sales and the payments to labor and capital is the nominal profit per year from running the family business. This nominal profit, which we represent by Π, is given by

profit = income from sales - wage and rental payments

$$\Pi = PY - (wL^d + RK^d).$$

If we substitute $A \cdot F(K^d, L^d)$ for Y, we get

(6.2) $$\Pi = PA \cdot F(K^d, L^d) - (wL^d + RK^d).$$

2. Wage income. If households supply the quantity of labor L^s to the labor market, they receive the nominal wage income of wL^s per year. As already mentioned, we assume that the quantity of labor supplied is the fixed amount L. Therefore, the nominal wage income is wL.

3. Rental income. If households supply the quantity of capital K^s to the rental market, they receive the nominal rental income of RK^s per year. Since households supply all of their available capital, K, to the rental market, so that $K^s = K$, the nominal rental income is RK.

We assume, as in chapter 3, that capital depreciates at the rate δ. Therefore, the quantity δK of capital disappears each year. The dollar value of this lost capital is $P\cdot\delta K$ per year. Hence, the net nominal rental income from ownership of capital is

$$\textit{net nominal rental income} = \textit{nominal rental income} - \textit{value of depreciation}$$

$$= RK - \delta PK.$$

If we divide and multiply the first term by P, we get

$$\textit{net nominal rental income} = (R/P)\cdot PK - \delta PK.$$

If we combine the terms involving PK, we get

$$(6.3) \qquad \textit{net nominal rental income} = (R/P - \delta)\cdot PK.$$

If we substitute $K = 1$ in equation (6.3), we see that the flow of net nominal rental income per unit of capital is $(R/P - \delta)\cdot P$ per year. That is, if a household pays $\$P$ to buy one unit of capital, it receives the flow of net nominal rental income of $(R/P - \delta)\cdot P$ per year. The rate of return per year to owning capital is the ratio of this net income to the amount paid per unit of capital, P:

$$\textit{rate of return to owning capital} = (\textit{net nominal rental income})/(\textit{nominal amount paid})$$

$$= (R/P - \delta)\cdot P / P.$$

Therefore, if we cancel out the P's, we get

$$(6.4) \qquad \textit{rate of return to owning capital} = R/P - \delta.$$

Hence, the rate of return per year on ownership of capital is the real rental price, R/P, less the rate of depreciation, δ.

4. Interest income. If a household's nominal bond holdings are B, the flow of nominal interest income received is iB per year. Notice that interest income is greater than zero for a holder of bonds (when B is greater than zero) and less than zero for an issuer of bonds (when B is less than zero). Since B equals zero for the whole economy, we have that the total of interest income equals zero. That is, the amount paid to holders of bonds (lenders) exactly balances the amount paid by issuers of bonds (borrowers).

5. Total income. We can put the four types of income together to calculate households' total nominal income per year. The result is

household nominal income = nominal profit + nominal wage income

+ nominal net rental income + nominal interest income.

If we substitute Π for nominal profit (from equation [6.2]), wL for nominal wage income, $(R/P-\delta)\cdot PK$ for nominal net rental income (from equation [6.3]), and iB for nominal interest income, we get

$$\text{(6.5)} \qquad \textit{household nominal income} = \Pi + wL + (R/P- \delta)\cdot PK + iB.$$

B. Consumption

So far, we have discussed household income. Now we consider household expenditures on goods. Households consume goods in the quantity C per year. Since the price level is P, the nominal amount spent on consumption per year is

$$\textit{household nominal consumption} = PC.$$

C. Assets

Now we work out how households' income and expenditure relate to households' assets. Households hold assets in three forms: money, M; bonds, B; and ownership of capital, K. Money pays no interest. Bonds pay interest at the rate i per year. Ownership of capital yields the rate of return $R/P - \delta$ per year. We assume that households can divide their assets any way they wish among the three forms. That is, at any point in time, households can exchange dollars of money for dollars of bonds and can exchange dollars of money for units of capital at the price level, P. So, when would households choose to hold all three forms of assets?

Bonds seem to be more attractive than money if the interest rate, i, is greater than zero. Households would, however, hold some money for convenience, because they use money to make various exchanges, including purchases and sales of goods and labor. In contrast, our assumption is that bonds are not readily accepted in exchange for goods or labor—usually, the holder of a bond has to sell the bond for money before buying goods or labor. The special role of money in making exchanges motivates households to have a positive **demand for money**. We postpone our study of this demand for money until chapter 10. For now, we just assume that households hold a fixed amount of money in dollar terms. That is, we assume that the change over time of a household's nominal money holdings is zero. If we use the symbol Δ to represent a change over time, we have

$$\Delta M = 0.$$

As an example, a family might want to hold \$200 on average to cover expenses for groceries, gasoline, and other goods. The amount of money held by an individual household would vary over time, sometimes rising above \$200 and sometimes falling below \$200. However, if the average money held by each household is always \$200, then the total money held at every point in time by all households would tend not to vary much. Our assumption is that the change in the total amount of money held by all households, ΔM, is zero.

If we consider bonds and capital, we would compare the rate of return to holding bonds—the interest rate, i—with the rate of return on ownership of capital—$R/P - \delta$. The question is whether households would be willing to hold both forms of assets if the rates of return differed. They might be willing to hold both if the assets differed by characteristics other than the rate of return. In the real world, the most important difference is the **riskiness** of the returns. For example, some types of bonds, such as U.S. Treasury Bills with 3-month maturity, are nearly **risk-free**.[3] Other forms of owning capital, such as stock in Ford Motor Company, provide returns (in the forms of dividends and capital gains) that are uncertain. In these situations, the risky asset (stock in Ford) typically has to pay an expected rate of return that is greater than the interest rate on U.S. Treasury Bills in order for people to be willing to hold the risky assets.

To keep things manageable in our model, we do not consider risk in the returns paid on bonds or capital. That is, we assume that, aside from rates of return, bonds and capital look the same to households as ways to hold assets. In this case, if bonds offered a higher rate of return, households would hold no capital. In contrast, if capital offered a higher rate of return, households would want to borrow a lot to hold a lot of capital (actually, an infinite amount). Since the economy's stock of capital is greater than zero but less than infinity, we must have that the two rates of return are equal. This condition is

---

[3] The holder of a three-month U.S. Treasury Bill has virtual certainty of receiving the promised dollar amount three months in the future. However, the real value of this payment is uncertain because the future price level is unknown, that is, the inflation rate is uncertain. Thus, U.S. Treasury Bills have risk in their real returns. Since 1997, the U.S. Treasury has issued inflation-protected securities (indexed bonds). These assets, if held to maturity, provide a guaranteed real return.

Key equation:

rate of return on bonds = rate of return on ownership of capital

(6.6) $$i = R/P - \delta.$$

If we use this result to substitute $R/P - \delta = i$ in the expression for household nominal income in equation (6.5), we get

(6.7) $$\text{household nominal income} = \pi + wL + i\cdot(B + PK).$$

The last term shows that assets held as bonds or ownership of capital yield the same rate of return per year, given by the interest rate, i.

Allowing for a risk premium on ownership of capital

We can readily extend the model to allow for a fixed **risk premium** on ownership of capital. In this case, the relation between the interest rate on bonds, i, and the rate of return on ownership of capital, $R/P - \delta$, would become

$$i = R/P - \delta - \text{risk premium}.$$

The risk premium is the excess of the anticipated rate of return on capital—for example, the expected rate of return on holding corporate stock—over the expected rate of return on a nearly risk-free asset, such as U.S. Treasury Bills. If the risk premium is constant, our analysis would not change. More interesting would be to allow for time-varying risk premia. Some reasons that risk premia can change are, first, the perceived riskiness of capital changes, second, households become more or less willing to absorb risk, and, third, innovations in the financial markets or the legal system make it easier for households to reduce the overall risk in their assets and incomes.

D. Household budget constraint

Now we use the results on household income to work out the household budget constraint. This constraint relates changes in households' assets to the flows of income.

At a point in time, a household has assets in the form of money, bonds, and ownership of capital:

$$\textit{nominal value of assets} = M + B + PK.$$

We define **nominal saving** to be the change over time in the nominal value of assets. Therefore, if we again use the symbol Δ to represent a change over time, we have

$$\begin{aligned} \textit{nominal saving} &= \Delta(\textit{nominal assets}) \\ &= \Delta M + \Delta B + P\cdot\Delta K. \end{aligned}$$

If we use our assumption that $\Delta M = 0$, we get

$$\textit{nominal saving} = \Delta B + P\cdot\Delta K. \tag{6.8}$$

That is, a household's saving corresponds to changes in its holdings of bonds or capital.

A household's nominal saving depends on its income and consumption. If income is greater than consumption, the difference will be saved and, therefore, added to assets. If income is less than consumption, nominal saving is less than zero, and the difference will subtract from nominal assets. Therefore, we have

$$\textit{nominal saving} = \textit{nominal income} - \textit{nominal consumption}.$$

If we substitute for nominal income from equation (6.7) and replace nominal consumption by PC, we get

$$(6.9) \qquad nominal\ saving = \Pi + wL + i\cdot(B + PK) - PC.$$

Equations (6.8) and (6.9) are two ways of representing nominal saving. Therefore, the right-hand sides of the equations must be equal:

$$(6.10) \qquad \Delta B + P\cdot\Delta K = \Pi + wL + i\cdot(B + PK) - PC.$$

This equation says that nominal saving, $\Delta B + P\cdot\Delta K$ on the left-hand side, equals the difference between nominal income, $\Pi + wL + i\cdot(B + PK)$, and nominal consumption, PC, on the right-hand side.

If we rearrange equation (6.10) to put nominal consumption, PC, on the left-hand side, we get

Key equation (household budget constraint):

$$(6.11) \qquad PC + \Delta B + P\cdot\Delta K = \Pi + wL + i\cdot(B + PK).$$

The right-hand side has total nominal income, $\Pi + wL + i\cdot(B + PK)$. Equation (6.11) says that households are constrained to divide this total nominal income between the two terms on the left-hand side: nominal consumption, PC, and nominal saving, $\Delta B + P\cdot\Delta K$. Thus, equation (6.11) is the **household budget constraint in nominal terms**.

We shall find it useful to express the household budget constraint in real terms, and we can do this by dividing everything in equation (6.11) by the price level, P. After we do this division, we get

Key equation (household budget constraint in real terms):

$$(6.12) \qquad C + (1/P)\cdot\Delta B + \Delta K = \Pi/P + (w/P)\cdot L + i\cdot(B/P + K).$$

This equation is the **household budget constraint in real terms**. The right-hand side has total real income, $\Pi/P + (w/P)\cdot L + i\cdot(B/P + K)$. The left-hand side has consumption, C, and the change in the real value of assets, $(1/P)\cdot\Delta B + \Delta K$. We refer to the change in the real value of assets as **real saving**. Notice that nominal saving, $\Delta B + P\cdot\Delta K$, which appears on the left-hand side of equation (6.11), gives the change in the nominal value of assets. In contrast, real saving, $(1/P)\cdot\Delta B + \Delta K$, which appears on the left-hand side of equation (6.12), gives the change in the real value of assets.

Figure 6.2 shows graphically the household budget constraint from equation (6.12). Suppose that a household has a given total real income, $\Pi/P + (w/P)\cdot L + i\cdot(B/P + K)$, on the right-hand side of the equation. The budget constraint says that this real income must be divided between consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$. One possibility is that the household sets real saving to zero, so that C equals the total real income. This choice corresponds to the point shown on the horizontal axis where C equals the total real income. Another possibility is that the household sets C to zero, so that real saving equals total real income. This choice corresponds to the point shown on the vertical axis where real saving equals total real income. More generally, the household can choose intermediate points where C and real saving are both greater than zero. The full range of possibilities is shown by the downward-sloping red line in the figure. This line is called a **budget line**. The budget constraint in equation (6.12) tells us that, along a budget line, each increase in C by one unit corresponds to a decrease in real saving by one unit. Hence, the slope of a budget line is -1.

V. Clearing of the Markets for Labor and Capital Services

Now that we have worked out the household budget constraint, we can consider the choices that households make. We begin by considering decisions about the family business. These decisions determine the demands for labor and capital services. Once we know these demands, we can study the clearing of the markets for labor and capital services.

A. Profit maximization

The two business decisions that households make are the quantities of labor and capital services to demand, L^d and K^d. These decisions determine the amount of goods produced and sold on the goods market, $A \cdot F(K^d, L^d)$, and therefore the amount of nominal profit from

(6.2) $$\Pi = PA \cdot F(K^d, L^d) - wL^d - RK^d.$$

To calculate real profit, we can divide through equation (6.2) by the price level, P, to get

(6.13) $$\Pi/P = A \cdot F(K^d, L^d) - (w/P) \cdot L^d - (R/P) \cdot K^d.$$

We can see from the right-hand side of the household budget constraint in equation (6.12) that an increase in real profit, Π/P, raises household real income. Figure 6.3 shows how an increase in real income affects households. An increase in real income moves the budget line outward from the red line to the blue one. In comparison with the red budget line, the blue line allows households to choose higher consumption, C, for any given value of real saving, $(1/P) \cdot \Delta B + \Delta K$. Therefore, as long as households like more consumption, they prefer more real income to less. This result tells us that households, as managers of family businesses, will seek to make real profit, Π/P, as high as possible. That is, households will choose their demands for labor and capital services, L^d and K^d, to maximize Π/P, as given in equation (6.13).

We assume that an individual household takes as given the real wage rate for labor, w/P, and the real rental price for capital, R/P. As mentioned before, these assumptions are standard for competitive markets—an individual household is too small to have a noticeable effect on the market prices. In this situation, each household can buy or sell whatever quantity of labor it wants at the going real wage rate, w/P, and can rent or rent out whatever amount of capital it wants at the going real rental price, R/P. Therefore, the household will demand quantities of labor and capital services, L^d and K^d, that maximize real profit, Π/P, for given values of w/P and R/P.

B. The labor market

We now consider the demand for labor and the supply of labor. Then we determine the real wage rate, w/P, from the market-clearing condition for the labor market: the quantity of labor demanded equals the quantity supplied.

1. Demand for labor. The demand for labor, L^d, comes from the objective of profit maximization. Consider the effect of an increase in labor input, L^d, by one unit on real profit, Π/P, as given in equation (6.13). The increase in L^d raises the first term on the right-hand side, $A \cdot F(K^d, L^d)$, by increasing production and, hence, sales of goods on the goods market. We know from chapter 3 that an increase in L^d by one unit raises production by the marginal product of labor, MPL. An increase in L^d also raises the second term in the equation—the real wage payments, $(w/P) \cdot L^d$. For a given real wage rate, w/P, an increase in L^d by one unit raises these payments by the amount w/P. Therefore, the overall effect from an increase in L^d by one unit is to change real profit by

$$\Delta(\Pi/P) = \Delta[A \cdot F(K^d, L^d)] - w/P$$

$$= MPL - w/P$$

change in real profit = marginal product of labor – real wage rate.

We know from chapter 3 that the marginal product of labor, MPL, depends on the quantity of labor input, L^d. As L^d rises, the MPL falls. This relation is shown in red by the downward-sloping curve in Figure 6.4. Note that this curve applies for a given technology level, A, and capital input, K^d.

Suppose that the household selects a low L^d, at which the MPL is greater than w/P in Figure 6.4. In this case, an increase in L^d by an additional unit would raise real profit, Π/P. The reason is that the addition to output—by MPL units—is larger than the addition to wage payments—by w/P units. However, as L^d rises, the MPL falls and eventually gets as low as w/P. If the household continues to raise L^d, the MPL falls below w/P. In that situation, further increases in L^d lower Π/P. Thus, to maximize real profit, the household should stop at the point where the MPL equals w/P. The graph shows that this equality occurs where the value along the curve for the MPL equals w/P.

For a given real wage rate, w/P, the graph in Figure 6.4 shows on the horizontal axis the quantity of labor demanded, L^d. We can see that a decrease in w/P raises L^d. Hence, if we graph L^d versus w/P, we map out a downward-sloping demand curve. That is, the labor-demand curve looks as shown in Figure 6.1.

Each household determines its labor demand, L^d, as shown in Figure 6.4. Therefore, when we add up across all the households, we end up with an aggregate or market demand for labor that also looks like the curve shown in the figure. In particular, a decrease in the real wage rate, w/P, raises the market quantity of labor demanded, L^d.

2. Supply of labor. We are assuming that each household supplies a fixed quantity of labor to the labor market. Therefore, the aggregate or market supply of labor, L^s, is the given amount L. More generally, the quantity of labor supplied would depend on the real wage rate, w/P. For example, we might have the upward-sloping labor-supply curve shown in Figure 6.1. However, we neglect for now this dependence of L^s on w/P.

3. Clearing of the labor market. The market labor demand, L^d, is determined from Figure 6.4 as a downward-sloping function of the real wage rate, w/P. We reproduce this curve in Figure 6.5. The market labor supply, L^s, is assumed to be the constant L. We show this fixed labor supply as the vertical line at L. Now we can determine the equilibrium value of w/P from the market-clearing condition for the labor market. Specifically, we assume that w/P is determined to equate the aggregate quantity of labor demanded, L^d, to the aggregate quantity supplied, L^s. This market-clearing value of w/P corresponds in Figure 6.5 to the intersection of the L^d curve with the vertical line at L. The market-clearing value for w/P is denoted by $(w/P)^*$. The corresponding market-clearing quantity of labor input is $L^* = L$.

The equality between L and L^d means that the market-clearing real wage rate, $(w/P)^*$, equals the marginal product of labor, MPL, that is,

$$(w/P)^* = MPL \text{ (evaluated at } L). \qquad (6.14)$$

By MPL (evaluated at L), we are referring in Figure 6.4 to the value for the marginal product of labor that corresponds to the quantity of labor L.[4]

Why do we assume that the labor market clears? The idea is that, only at this market-clearing position, would the real wage rate, w/P, tend neither to rise nor fall. If w/P were very low, the aggregate quantity of labor demanded would exceed the quantity supplied. In this case, demanders of labor would compete to hire scarce workers by raising w/P.[5] Conversely, if the quantity demanded were less than the quantity supplied, w/P would be bid down. In equilibrium, w/P will be determined to clear the labor market, that is, so that the aggregate quantity of labor supplied, $L^s = L$, equals the quantity demanded, L^d.

C. The market for capital services

We now consider the demand for capital services and the supply of capital services. Then we determine the real rental price, R/P, from the market-clearing condition for the market for capital services: the quantity of capital services demanded equals the quantity supplied.

1. Demand for capital services. As with the demand for labor, the demand for capital services, K^d, comes from the objective of profit maximization. Consider the effect of an increase in K^d by one unit on a household's real profit, as given again by

$$\Pi/P = A \cdot F(K^d, L^d) - (w/P) \cdot L^d - (R/P) \cdot K^d. \qquad (6.13)$$

[4] Note that the curve for MPL in Figure 6.4 applies for a given capital stock, K. A change in K would shift the MPL associated with a given value of L and would therefore change $(w/P)^*$.

[5] Our description views the participants in the labor market as directly setting the real wage rate, w/P. However, our story would be the same if, instead, the price level, P, were given, and the participants in the labor market adjusted the nominal wage rate, w.

We know from chapter 3 that an increase in capital input, K^d, by one unit raises production, $A \cdot F(K^d, L^d)$, by the marginal product of capital, MPK. An increase in K^d also raises the last term in the equation—the real rental payments, $(R/P) \cdot K^d$. For a given real rental price, R/P, an increase in K^d by one unit raises these payments by the amount R/P. Therefore, the overall effect from an increase in K^d by one unit is to change real profit by

$$\Delta(\Pi/P) = \Delta[A \cdot F(K^d, L^d)] - R/P$$

$$= MPK - R/P$$

change in real profit = marginal product of capital – real rental price.

We know from chapter 3 that the marginal product of capital, MPK, depends on the amount of capital input, K^d. As K^d rises, MPK falls. This relation is shown in red by the downward-sloping curve in Figure 6.6. Note that this curve applies for a given technology level, A, and labor input, L^d.

Suppose that the household selects a low K^d, at which the MPK is greater than R/P in Figure 6.6. In this case, an increase in K^d by one unit would raise real profit, Π/P. The reason is that the addition to output—by MPK units—is greater than the increase in real rental payments—by R/P units. However, as K^d rises, the MPK falls and eventually gets as low as R/P. If the household continues to raise K^d, the MPK falls below R/P. In that case, further increases in K^d lower Π/P. Thus, to maximize real profit, the household should stop at the point where the MPK equals R/P. The graph shows that this equality occurs where the value along the curve for the MPK equals R/P.

For a given real rental price, R/P, the graph in Figure 6.6 shows on the horizontal axis the quantity of capital services demanded, K^d. We can see that a decrease in R/P

increases K^d. Hence, if we graph K^d versus *R/P*, we map out a downward-sloping demand curve.

Each household in the economy determines its demand for capital services, K^d, as shown in Figure 6.6. Therefore, when we add up across all the households, we end up with an aggregate or market demand for capital services that also looks like the curve shown in the figure. In particular, a decrease in the real rental price, *R/P*, raises the market quantity of capital services demanded, K^d.

2. Supply of capital services. For the economy as a whole, the aggregate quantity of capital, *K*, is given from past decisions on investment. That is, in the short run, the economy has a given stock of houses, cars, machines, and factories. This capital stock is owned by households, and all of the services from this stock are supplied to the rental market. Therefore, the aggregate or market quantity capital services supplied is $K^s = K$.

3. Clearing of the market for capital services. The market demand for capital services, K^d, is determined from Figure 6.6 as a downward-sloping function of the real rental price, *R/P*. We reproduce this curve in Figure 6.7. The market supply of capital services, K^s, is the constant *K*. We show this fixed supply of capital services as the vertical line at *K*. As with the labor market, we assume that the equilibrium value of *R/P* is determined to clear the market, that is, so that the aggregate quantity of capital services supplied, $K^s = K$, equals the aggregate quantity demanded, K^d. This market-clearing value of *R/P* corresponds in Figure 6.7 to the intersection of the K^d curve with the vertical line at K. The market-clearing value for R/P is denoted by $(R/P)^*$. The corresponding market-clearing quantity of capital services is K.

The equality between K and K^d means that the market-clearing real rental price, $(R/P)^*$, equals the marginal product of capital, MPK, that is,

(6.15) $$(R/P)^* = MPK \text{ (evaluated at } K).$$

By MPK (evaluated at K), we are referring in Figure 6.6 to the value for the marginal product of capital that corresponds to the capital input K.[6]

We can ask again why we assume that the market clears. The idea is that, only at this market-clearing position, would the real rental price, R/P, tend neither to rise nor fall. If R/P were very low, the aggregate quantity of capital services demanded would exceed the quantity supplied. In this case, demanders of capital services would compete to hire scarce capital by raising R/P. Conversely, if the quantity demanded were less than the quantity supplied, R/P would be bid down. In equilibrium, R/P will be determined to clear the market, that is, so that the aggregate quantity of capital services supplied, $K^s = K$, equals the aggregate quantity demanded, K^d.

4. The interest rate. The market-clearing solution for the real rental price, $R/P = (R/P)^*$, will allow us to determine the interest rate, i. We found before that i, which is the rate of return on bonds, equals the rate of return to owning capital:

(6.4) $$i = R/P - \delta.$$

Therefore, if we substitute for R/P from the formula for $(R/P)^*$ in equation (6.15), we get

(6.16) $$i = MPK \text{ (evaluated at } K) - \delta.$$

---

[6] Note that the curve for MPK in Figure 6.6 applies for a given labor input, L. A change in L would shift the MPK associated with a given value of K and would therefore change $(R/P)^*$.

Thus, once we determine R/P, we also determine the interest rate, i.

This last result is important. It says that the interest rate, i, cannot change unless something changes the marginal product of capital, MPK. For a given technology level, A, the MPK depends on the inputs of capital services, K, and labor, L. In our present setting, K and L are given. Therefore, i will also be given. In the real world, the interest rate, i, tends to fluctuate a good deal. Thus, to capture this aspect of reality, we shall have to extend our model to allow for variations in the MPK. We introduce in the next two chapters some sources of changes in the MPK and, therefore, in the interest rate, i.

D. Profit in equilibrium

We determined households' demands for labor and capital services, L^d and K^d, from the objective of profit maximization. That is, households chose L^d and K^d to make real profit, Π/P, as high as possible. Now we consider the level of Π/P that households receive when the labor and rental markets clear.

When the labor and rental markets clear, so that $L^d = L$ and $K^d = K$, real profit is given from equation (6.13) by

$$\Pi/P = A \cdot F(K,L) - (w/P) \cdot L - (R/P) \cdot K. \tag{6.17}$$

Since the labor and rental markets clear, we also have, from equations (6.14) and (6.15),

$$w/P = MPL,$$

$$R/P = MPK,$$

where MPL and MPK are evaluated at the given values of L and K. If we substitute these formulas for w/P and R/P into equation (6.17), we get

(6.18) $$\Pi/P = A \cdot F(K,L) - MPL \cdot L - MPK \cdot K.$$

We show in the appendix that the expression on the right-hand side of equation (6.18) equals zero. That is, when the values for w/P and R/P satisfy the market-clearing conditions for labor and capital services, the real profit, Π/P, ends up being zero. Another way to say this is that real GDP, which equals $A \cdot F(K,L)$, just covers the payments to the two factor inputs, $(w/P) \cdot L$ for labor and $(R/P) \cdot K$ for capital services. Output equals the total of these factor incomes, and all of this income goes to either labor or capital. Nothing is left over for profit.

We therefore have something of a paradox. Households choose demands for labor and capital services, L^d and K^d, to maximize profit. However, when w/P and R/P satisfy market-clearing conditions, the resulting real profit, Π/P, is zero. That is, the highest real profit that households can attain is zero.

Economic profit versus accounting profit

Our definition of profit differs from the standard accounting definition. The reason for the difference is the treatment of rental payments on capital. Suppose, for example, that a household (or, more realistically, a business) owns capital and uses the capital to produce goods. In that case, the household pays no explicit rental payments on the capital that it uses in production. The rental payments are only implicit—the household should think of paying rentals to itself on the capital that it owns and uses. These implicit rental payments represent the income that the household could have received by renting the capital out to another producer. Thus, the foregone rental payments should be treated as a cost (called an opportunity cost) of using one's own capital to produce goods.

Standard accounting practices, including the national-income accounts, do not include most forms of implicit rental payments as costs.[7] For this reason, the usual accounting measure of rental payments understates the rental payments that are appropriate from an economic perspective. Since rental payments are a negative item for real profit, *Π/P*, in equation (6.17), we also have that the accounting measure overstates real profit from an economic standpoint. Since the economic profit is zero in equilibrium, we therefore have that the accounting measure of profit is greater than zero in equilibrium. The accounting measure of profit really measures the uncounted part of the rental income on capital.

VI. Summing Up

We set up the market structure and microeconomic foundations for our macroeconomic model. We began by describing the economy's four markets—for goods, labor, rental of capital, and bonds. The dollar prices on the first three markets are

[7] If a business borrows money to finance purchases of capital goods, the usual accounting practice treats the interest payments on the business's debt as costs. Therefore, interest expenses—which represent the rental payments on debt-financed capital—enter as a negative item in the accounting definition of profit. Thus, standard accounting procedures include rental payments on capital as costs when the capital is debt financed but not when the capital is owned outright by a business. In the case of corporations, this ownership corresponds to finance of capital by issue of equity shares (corporate stock) or retained earnings.

the price level, P, the wage rate, w, and the rental price, R. The price of short-term bonds is fixed at one, but the interest rate, i, can vary. We can think of i as the price of credit. We showed how the sales and purchases on the various markets determined household income from wages, returns on assets, and profit.

We examined in detail the markets for labor and capital services. As managers of family businesses, households determined demands for labor, L^d, and capital services, K^d. We derived these demands from the objective of profit maximization, assuming that each household took as given the real wage rate, w/P, and the real rental price, R/P. We assumed that the supply of labor, L^s, was fixed at L and that the supply of capital services, K^s, was fixed at K.

We found from the labor market that the market-clearing real wage rate, $(w/P)^*$, equaled the marginal product of labor, MPL. We found from the market for capital services that the market-clearing real rental price, $(R/P)^*$, equaled the marginal product of capital, MPK. The two marginal products, MPL and MPK, were evaluated at the given values of L and K. We also showed that the market-clearing real rental price, $(R/P)^*$, determined the interest rate, i.

Questions and Problems

Mainly for review

6.1. Why would households be interested only in the real values of consumption, income, and assets such as bonds? Think about how households would feel if the nominal values of consumption, income, and assets all double and the price level, P, also doubles.

6.2. Distinguish clearly between a household's initial asset position and the change in that position. If a household has negative saving, is that household necessarily a borrower in the sense of having a negative position in bonds?

6.3. Derive the budget line shown in Figure 6.3. What does this line show?

6.4. How does an increase in the real wage rate, w/P, affect the quantity of labor demanded, L^d? Where does the assumption of diminishing marginal productivity of labor come in?

6.5. How does an increase in the real rental price, R/P, affect the quantity of capital services demanded, K^d? Where does the assumption of diminishing marginal productivity of capital come in?

Problems for discussion

6.x. Discount bonds

The bonds in our model have a maturity close to zero; they just pay the current interest rate, *i*, as a flow over time. We could consider, instead, a discount bond, such as a U.S. Treasury Bill. This type of asset has no explicit interest payments (called coupons) but pays a principal of, say, \$1000 at a fixed date in the future. A Bill with one-year maturity pays off one year from the issue date, and similarly for 3-month or 6-month Bills. Let P^B be the price of a discount bond with one-year maturity and principal of \$1000.

a. Is P^B greater than or less than \$1000?

b. What is the one-year interest rate on these discount bonds?

c. If P^B rises, what happens to the interest rate on these bonds?

d. Suppose that, instead of paying \$1000 in one year, the bond pays \$1000 in two years? What is the interest rate per year on this two-year bond?

6.x. Term structure of interest rates

Suppose that the economy has discount bonds (discussed in equation 6.x) with one- and two-year maturity. Let i_t^1 be the interest rate on a one-year bond issued at the start of year *t* and i_{t+1}^1 the interest rate on a one-year bond issued at the start of year *t+1*. Let i_t^2 be the interest rate on a two-year bond issued at the start of year *t*. We can think of i_t^1 as the current short-term interest rate and i_t^2 as the current long-term interest rate.

a. Assume that, at the start of year *t*, everyone knows not only i_t^1 and i_t^2, but also the next year's one-year (short-term) rate, i_{t+1}^1. What must be the relation of i_t^2 to i_t^1 and i_{t+1}^1? Explain the answer by considering the incentives of lenders and borrowers.

b. If $i_{t+1}^1 > i_t^1$, what is the relation between i_t^2, the long-term interest rate, and i_t^1, the short-term interest rate? The answer is an important result about the term structure of interest rates.

c. How would the results be affected if we assumed, more realistically, that there was uncertainty in year *t* about the future one-year interest rate, i_{t+1}^1?

6.x. Financial intermediaries

Consider a financial intermediary, such as a bank, that participates in the credit market. This intermediary borrows from some households and lends to others. (The loan from a customer to a bank often takes the form of a deposit account.)

a. Does the existence of intermediaries affect the result that the aggregate amount of loans is zero?

b. What interest rates would the intermediary charge to its borrowers and pay to its lenders? Why must there be some spread between these two rates?

c. Can you give some reasons to explain why intermediaries might be useful?

Appendix: Output equals factor incomes and profit equals zero

We show here that, when capital and labor are each paid their marginal products, the total of income payments to capital and labor equals the output or real GDP. Hence, profit is zero. These results are shown most easily using calculus.

Start with the production function for real GDP, Y:

(3.1) $$Y = A \cdot F(K,L).$$

In chapter 3, we assumed that the production function satisfied constant returns to scale in capital and labor, K and L. Therefore, if we multiply K and L each by $1/L$, we also multiply Y by $1/L$:

$$Y/L = A \cdot F(K/L,\ 1).$$

Thus, output per unit of labor, Y/L, depends only on capital per unit of labor, K/L. If we multiply through each side of the equation by L, we get another way to write the production function:

(6.19) $$Y = AL \cdot F(K/L,\ 1).$$

We can use calculus to calculate the marginal product of capital, MPK, from equation (6.19). The MPK is the derivative of Y with respect to K, while holding fixed A and L:

$$MPK = AL \cdot F_1 \cdot (1/L),$$

where F_1 is the derivative of the function F with respect to its first argument, K/L. We get the last term, $1/L$, from the chain rule for differentiation. That is, $1/L$ is the derivative of K/L with respect to K, while holding fixed L. If we cancel out L and $(1/L)$, we get

(6.20) $$MPK = AF_1.$$

The marginal product of labor, MPL, is the derivative of Y with respect to L, while holding fixed A and K. Since L appears in two places on the right-hand side of equation (6.19), we have to calculate the derivative with respect to L of the product of two terms, AL and $F(K/L, 1)$. The first part of the answer is the derivative of the first term, A, multiplied by the second term, $F(K/L, 1)$. The second part is the first term, AL, multiplied by the derivative of the second term. This derivative is $F_1 \cdot (-K/L^2)$, where F_1 is again the derivative of the function F with respect to its first argument, K/L. The term $-K/L^2$ comes from the chain rule for differentiation. That is, $-K/L^2$ is the derivative of K/L with respect to L, while holding fixed K. Putting the results together, we get

$$MPL = A \cdot F(K/L, 1) + A\cancel{L} \cdot F_1 \cdot (-K/L^{\cancel{2}})$$

$$MPL = A \cdot F(K/L, 1) - A \cdot (K/L) \cdot F_1. \tag{6.21}$$

If the factor inputs are each paid their marginal products, so that $w/P = MPL$ and $R/P = MPK$, we can use equations (6.20) and (6.21) to calculate the total payments to labor and capital:

$$\begin{aligned}(w/P) \cdot L + (R/P) \cdot K &= MPL \cdot L + MPK \cdot K \\ &= [A \cdot F(K/L, 1) - A \cdot (K/L) \cdot F_1] \cdot L + (AF_1) \cdot K \\ &= AL \cdot F(K/L, 1) - \cancel{AK \cdot F_1} + \cancel{AK \cdot F_1} \\ &= AL \cdot F(K/L, 1).\end{aligned}$$

Equation (6.19) tells us that the last term equals Y, which equals $A \cdot F(K,L)$. Therefore, we have shown:

$$(w/P) \cdot L + (R/P) \cdot K = A \cdot F(K,L). \tag{6.22}$$

Thus, the total payments to labor and capital, $(w/P)\cdot L + (R/P)\cdot K$, equal real GDP, $A\cdot F(K,L)$.[8]

Recall that real profit is given by

(6.17) $$\Pi/P = A\cdot F(K, L) - (w/P)\cdot L - (R/P)\cdot K.$$

Therefore, the result in equation (6.22) proves that Π/P is zero in equilibrium, as claimed in the text.

[8] This result is called Euler's Theorem.

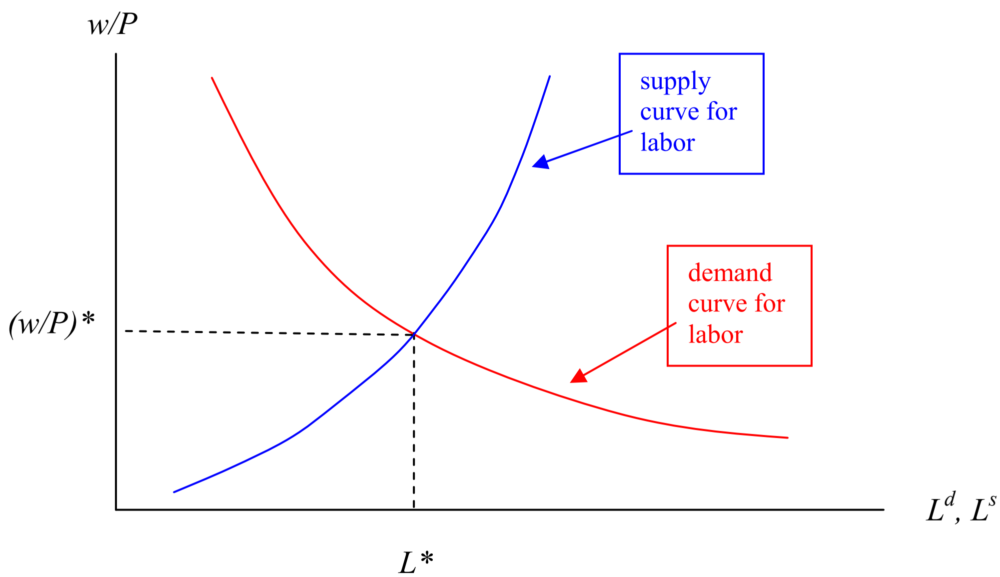


Figure 6.1

An Example of Market Clearing: The Labor Market

This figure gives a simple example of how a market—in this case, the labor market—would clear. The labor-demand curve, L^d, slopes downward versus the real wage rate, w/P. The labor-supply curve, L^s, slopes upward versus w/P. Market clearing corresponds to the intersection of the two curves. The market-clearing real wage rate is $(w/P)^*$, and the market-clearing quantity of labor is L^*.

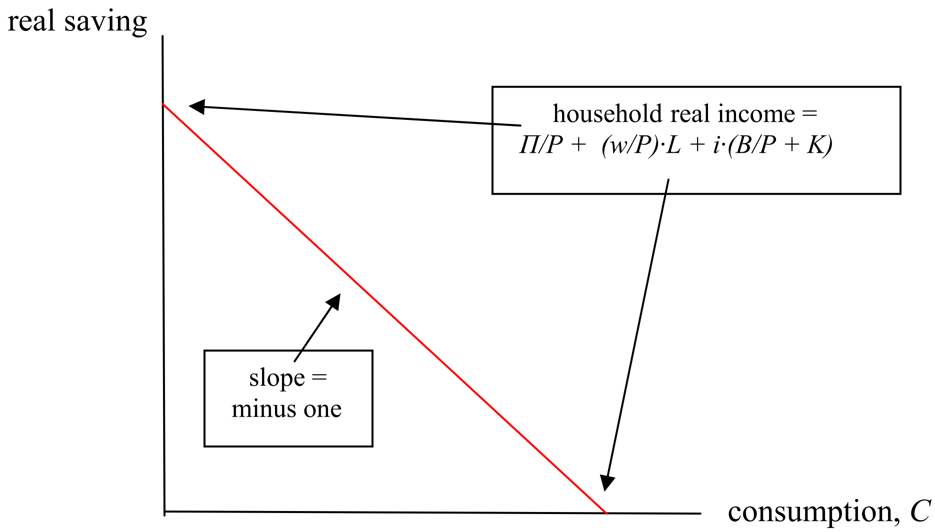


Figure 6.2

The Household Budget Constraint

Households have a given total of real income, $\Pi/P + (w/P)\cdot L + i\cdot(B/P + K)$. This total must be divided between consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$. Thus, if real saving is zero, C equals the total of real income along the horizontal axis. If C is zero, real saving equals the total of real income along the vertical axis. The budget constraint in equation (6.12) allows the household to select any combination of consumption and real saving along the red line, which has a slope of minus one. Along this budget line, one unit less of consumption corresponds to one unit more of real saving.

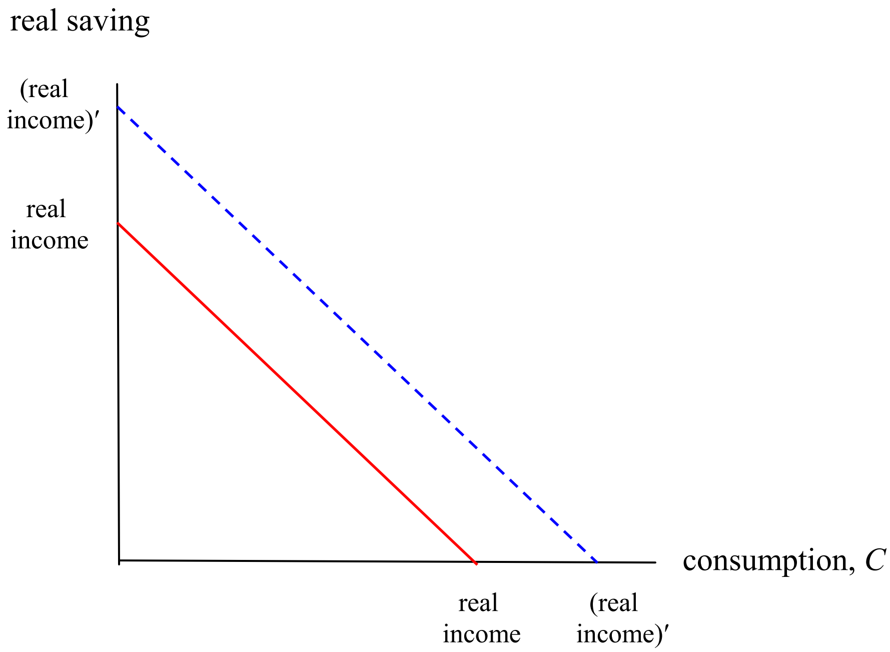


Figure 6.3

Effect of an Increase in Real Income on the Household Budget Constraint

If household real income, $\Pi/P + (w/P)\cdot L + i\cdot(B/P + K)$, rises, the budget line moves outward from the red line to the blue line. In comparison with the red line, the blue line allows the households to have more consumption, C, for any given value of real saving, $(1/P)\cdot\Delta B + \Delta K$. Thus, if households like more consumption, they prefer more real income to less.

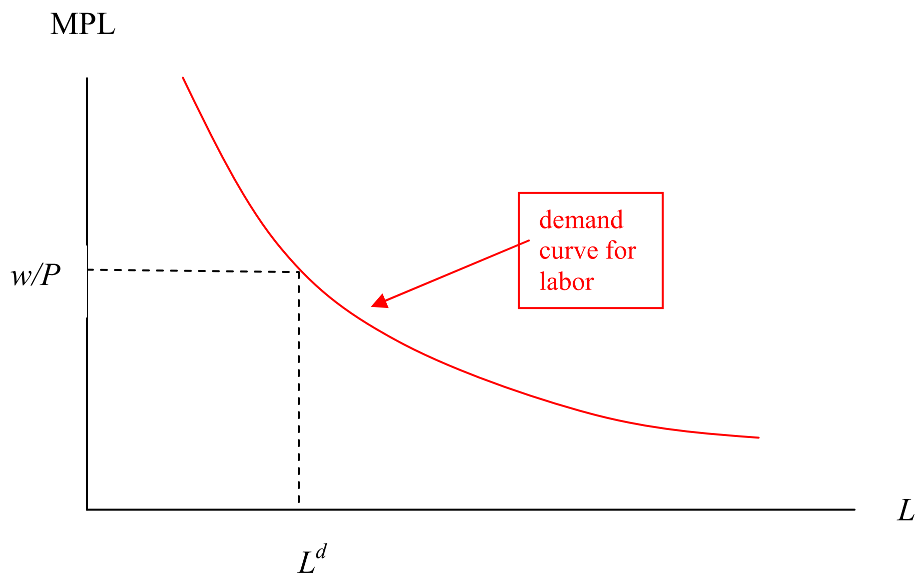


Figure 6.4

Labor Demand

For a given technology level, A, and capital input, K^d, the marginal product of labor, MPL, decreases as labor input, L, increases. Therefore, the MPL declines on the vertical axis as L rises on the horizontal axis. The household chooses labor input, L^d, where the MPL equals the real wage rate, w/P. If w/P decreases, L^d increases. Therefore, the labor-demand curve is downward sloping, as assumed in Figure 6.1.

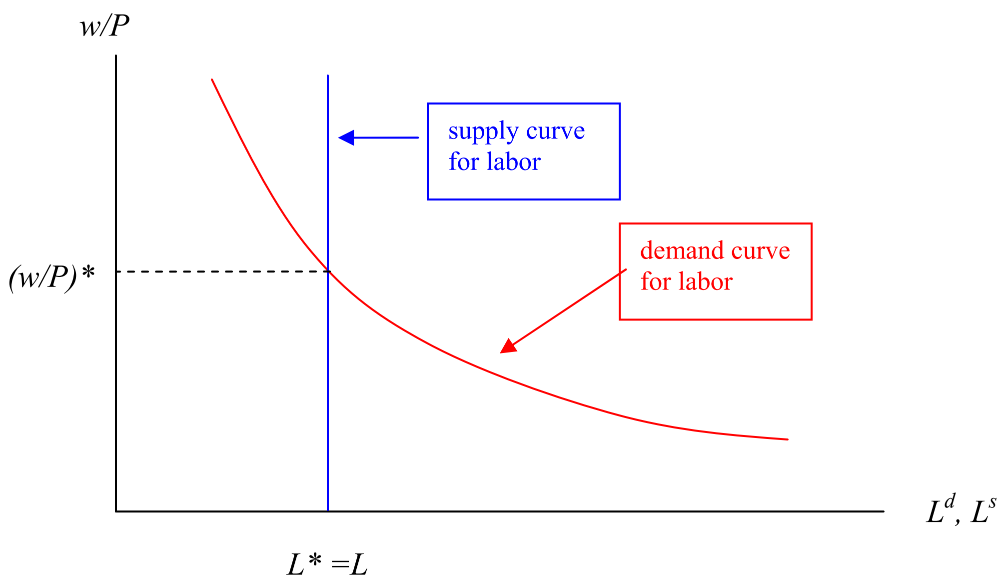


Figure 6.5

Clearing of the Labor Market

The downward-sloping labor-demand curve, L^d, comes from Figure 6.4. We assume here that labor supply, L^s, is fixed at L. The market-clearing real wage rate is $(w/P)^*$. The market-clearing quantity of labor input is $L^* = L$.

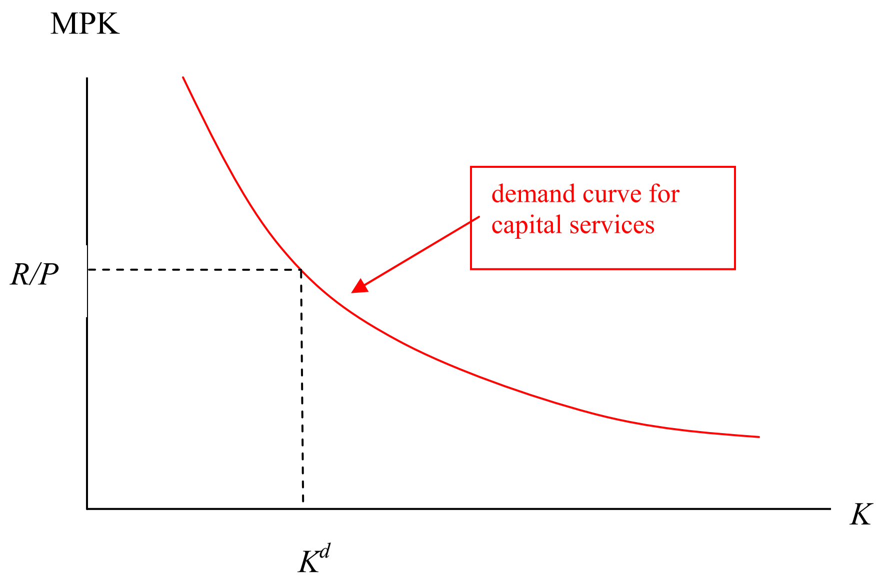


Figure 6.6

Demand for Capital Services

For a given technology level, A, and labor input, L^d, the marginal product of capital, MPK, decreases as capital input, K, increases. Therefore, the MPK declines on the vertical axis as K rises on the horizontal axis. The household chooses capital input, K^d, where the MPK equals the real rental price, R/P. If R/P decreases, K^d increases. Therefore, the demand curve for capital services is downward sloping.

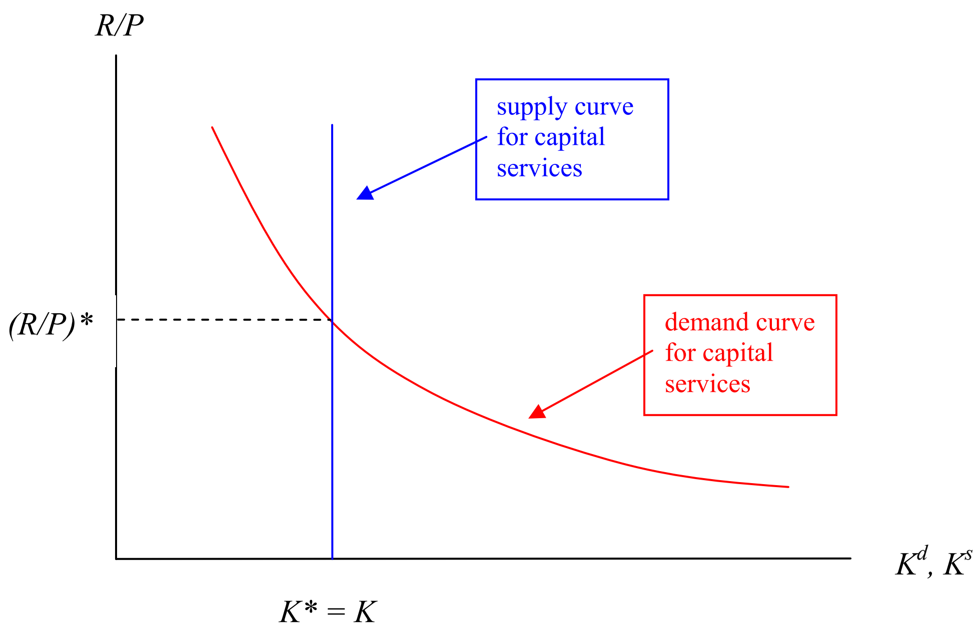


Figure 6.7

Clearing of the Market for Capital Services

The downward-sloping demand curve for capital services, K^d, comes from Figure 6.6. The supply of capital services, K^s, is fixed at K. The market-clearing real rental price is $(R/P)^*$. The market-clearing quantity of capital services is $K^* = K$.

January 12, 2005

Chapter 7

Consumption, Saving, and Real Business Cycles

In the previous chapter, we discussed the four markets in our model of the macro economy—goods, labor, capital services, and bonds. We studied how household income related to the prices and quantities in the four markets. We began the construction of the model's microeconomic foundations by considering how households, as managers of family businesses, determined their demands for labor and capital services. Then we investigated the clearing of the markets for labor and capital services. For given values of labor, L, and capital, K, these market-clearing conditions determined the real wage rate, w/P, the real rental price of capital, R/P, and the interest rate, i.

In the first part of this chapter, we extend our microeconomic analysis of households to the choices of consumption and saving. We use these results to determine economy-wide levels of consumption, saving, and investment. Then we use the analysis from this and the preceding chapter to study the economic fluctuations that result from shifts to the technology level, A. This kind of analysis is called **real business cycle theory**. We use this theory to study fluctuations in real GDP and other macroeconomic variables, including consumption, saving, investment, the real wage rate, the real rental price of capital, and the interest rate.

Later in this chapter we look in detail at the U.S. data to see how macroeconomic variables typically behave during economic fluctuations. However, to motivate our theoretical analysis, we summarize some of the key facts here. First, consumption and investment typically move in the same direction as real GDP—that is, both are high in booms and low in recessions. Second, consumption is less variable than real GDP, whereas investment is far more variable than real GDP. Third, the real wage rate, the real rental price of capital, and (less clearly) interest rates tend to move in the same direction as real GDP—high in booms, low in recessions. These are some of the facts that we want our model to explain.

I. Consumption and Saving

Now we study households' choices of consumption, C, which will determine real saving. Start with the household budget constraint from equation (6.12) of chapter 6:

$$C + (1/P)\cdot\Delta B + \Delta K = \Pi/P + (w/P)\cdot L + i\cdot(B/P + K). \quad (6.12)$$

We showed in chapter 6 that real profit, Π/P, equaled zero when the markets for labor and capital services cleared. Therefore, we can set $\Pi/P = 0$ to get a simplified form of the household budget constraint:

$$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L + i\cdot(B/P + K). \quad (7.1)$$

Hence, consumption, C, plus real saving, $(1/P)\cdot\Delta B + \Delta K$, equals real wage income, $(w/P)\cdot L$, plus real income on assets, $i\cdot(B/P + K)$.

We assumed before that an individual household takes as given the real wage rate, w/P. This assumption is a standard one for a competitive market—the individual household is too small to have a noticeable effect on the real wage rate. Now we assume that an individual household also takes as given the interest rate, i. This assumption is another standard one for a competitive market—the individual household is too small to have a noticeable effect on the interest rate. In this setup, each household can lend or borrow whatever amount it wants at the going interest rate. A household would borrow by issuing a bond that pays the interest rate, i. A household would lend by buying a bond that pays the interest rate, i.

Suppose that a household has given labor, L, and real assets, $(B/P + K)$. Then, since the real wage rate, w/P, and the interest rate, i, are given by the market to an individual household, the total of household real income, $(w/P)\cdot L + i\cdot(B/P + K)$, is given on the right-hand side of equation (7.1). The household can then choose how to divide this real income between consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$, on the left-hand side of the equation. That is, the household budget constraint pins down the total of consumption and real saving. If the household wants one unit more of consumption it must give up one unit of real saving.

Figure 7.1 (which comes from Figure 6.3) shows how the budget constraint from equation (7.1) allows the household to choose between consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$. If real saving were set to zero, all of the real income, $(w/P)\cdot L + i\cdot(B/P + K)$, would go to consumption. Hence, C would equal this value along the horizontal axis. Alternatively, if C were set to zero, all of the real income would go to real saving. Hence, real saving would equal the real income, $(w/P)\cdot L + i\cdot(B/P + K)$, along the vertical axis. More generally, the household can take a less extreme approach by not allocating its full income to consumption on the one hand or real saving on the other hand. That is, the household can choose an intermediate position where consumption and real saving are both greater than zero. The red line, which has a slope of minus one, shows the possibilities. Along this line, if C falls by one unit, real saving rises by one unit.

We want to consider a household's choices of consumption at different points in time. We know that more consumption today means less real saving. The decrease in real saving leads to lower real assets in the future. These assets are a source of future income. Therefore, with fewer real assets available, consumption has to be lower sometime in the future. Thus, the basic tradeoff is that more consumption today means less consumption later.

To work out formally the choices of consumption over time, we have to consider that the household budget constraint, equation (7.1), applies at every point in time. The link between today's budget constraint and tomorrow's comes from the effect of today's real saving, $(1/P)\cdot\Delta B + \Delta K$, on tomorrow's real assets, $B/P + K$. We can go a long way in understanding this linkage by considering just two periods of discrete length. To be concrete, we think of the first period as the current year and the second period as the next year.

A. Consumption over two years

For the current year, year 1, the budget constraint from equation (7.1) translates into

$$(7.2) \qquad C_1 + (B_1/P + K_1) - (B_0/P + K_0) = (w/P)_1 \cdot L + i_0 \cdot (B_0/P + K_0).$$

On the left-hand side, C_1 is year 1's consumption. The real assets B_1/P and K_1 are the amounts held at the end of year 1. The real assets B_0/P and K_0 are the amounts held at the end of the previous year, year 0, and therefore also the amounts held at the *beginning* of year 1. Thus, $(B_1/P + K_1) – (B_0/P + K_0)$ is the change in real assets—or real saving—during year 1.

On the right-hand side, the real wage rate for year 1 is $(w/P)_1$, and the real wage income for the year is $(w/P)_1 \cdot L$. For now, we continue to assume that labor, L, is constant over time. Therefore, we do not include a year subscript. The interest rate applicable to assets held at the end of the previous year is i_0. Thus, the real interest income for year 1 is $i_0 \cdot (B_0/P + K_0)$.

The budget constraint in equation (7.2) is for year 1. However, the same form applies to year 2:

$$(7.3) \qquad C_2 + (B_2/P + K_2) - (B_1/P + K_1) = (w/P)_2 \cdot L + i_1 \cdot (B_1/P + K_1).$$

What do the budget constraints in equations (7.2) and (7.3) tell us about households' choices between consuming this year, C_1, and next year, C_2? To understand this choice, notice that the two budget constraints are connected because both include the assets held at the end of year 1, $(B_1/P + K_1)$. In equation (7.2), if we move C_1 and $(B_0/P + K_0)$ from the left-hand side to the right-hand side and combine the two terms that have $(B_0/P + K_0)$, we find that the assets at the end of year 1 are

$$(7.4) \qquad (B_1/P + K_1) = (1+i_0) \cdot (B_0/P + K_0) + (w/P)_1 \cdot L - C_1.$$

Thus, if the household lowers year 1's consumption, C_1, by one unit, the assets held at the end of year 1, $(B_1/P + K_1)$ rise by one unit.

Similarly, equation (7.3) implies that the assets held at the end of year 2 are given by

$$(7.5) \qquad (B_2/P + K_2) = (1+i_1) \cdot (B_1/P + K_1) + (w/P)_2 \cdot L - C_2.$$

On the right-hand side, an increase in $(B_1/P + K_1)$—the assets held at the end of year 1—allows households to raise year 2's consumption, C_2, without changing the assets held at the end of year 2, $(B_2/P + K_2)$. If we put this result together with the one in equation (7.4), we find that a reduction in this year's consumption, C_1, allows the household to increase $(B_1/P + K_1)$ and, thereby, next year's consumption, C_2. Moreover, the household can switch between C_1 and C_2 without changing $(B_2/P + K_2)$, the real assets carried over to year 3. Thus, the household is changing the timing of consumption between this year and next year, C_1 and C_2, without short-changing or enriching the future, that is, without changing the assets available for year 3.

We can find the precise connection between C_1 and C_2 by replacing $(B_1/P + K_1)$ on the right-hand side of equation (7.5) by the right-hand side of equation (7.4) to get

$$(B_2/P + K_2) = (1+i_1)\cdot[(1+i_0)\cdot(B_0/P + K_0) + (w/P)_1\cdot L - C_1] + (w/P)_2\cdot L - C_2.$$

If we multiply through $(1+i_1)$ by the terms inside the brackets, we get

$$(7.6)\quad (B_2/P + K_2) = (1+i_1)\cdot(1+i_0)\cdot(B_0/P + K_0) + (1+i_1)\cdot (w/P)_1\cdot L - (1+i_1)\cdot C_1 + (w/P)_2\cdot L - C_2.$$

Consider the way that the interest rates enter on the right-hand side of equation (7.6). The first term has the assets from the end of year 0, $(B_0/P + K_0)$, which pay the return $i_0\cdot(B_0/P + K_0)$ in year 1. If the household holds these assets, it ends up with $(1+i_0)\cdot(B_0/P + K_0)$ in assets at the end of year 1. Hence, each unit of assets from the end of year 0 is multiplied by $1+i_0$ to get assets at the end of year 1.

The same calculation applies for year 2. Each unit of assets held at the end of year 1 is multiplied by $1+i_1$ to get assets at the end of year 2. If we put this result together with the one for year 1, we find that each unit of assets held from the end of year 0 to the end of year 2 ends up as $(1+i_1)\cdot(1+i_0)$ units of assets. This interest-rate term is the one that multiplies $(B_0/P + K_0)$ on the right-hand side of equation (7.6).

Similarly, the wage income $(w/P)_1\cdot L$ could be used to hold assets at the end of year 1, and each unit of these assets becomes $1+i_1$ units of assets at the end of year 2. Therefore, $(w/P)_1\cdot L$ is multiplied by $1+i_1$ in equation (7.6). In contrast, the wage income $(w/P)_2\cdot L$ appears by itself because this income comes too late to yield any asset returns in year 2. The consumption terms enter like the income terms, except for the signs. Therefore, C_1, like $(w/P)_1\cdot L$, is multiplied by $1+i_1$, and C_2, like $(w/P)_2\cdot L$, appears by itself.

If we divide through everything in equation (7.6) by $(1+i_1)$ and rearrange the terms to put the ones involving consumption on the left-hand side, we get

Key equation (household budget constraint for two years):

$$(7.7) \quad C_1 + C_2/(1+i_1) = (1+i_0)\cdot(B_0/P+K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L\,/(1+i_1) - (B_2/P+K_2)/(1+i_1).$$

Recall that we got this result by using the budget constraints for years 1 and 2, as given in equations (7.2) and (7.3). Therefore, we call equation (7.7) the **two-year budget constraint**.

Observe how the wage incomes, $(w/P)_1\cdot L$ and $(w/P)_2\cdot L$, enter on the right-hand side of equation (7.7). We do not add the two together, but rather divide $(w/P)_2\cdot L$ by $1+i_1$ before combining it with $(w/P)_1\cdot L$. Similarly, on the left-hand side, we make the same adjustment to C_2 before combining it with C_1. To understand these adjustments, we have to consider **present values**.

1. Present values. If the interest rate, i_1, is greater than zero, one dollar held as assets in year 1 becomes more than one dollar in year 2. Therefore, \$1 received or spent in year 1 is equivalent to more than \$1 in year 2. Or, viewed in reverse, dollars received or spent in year 2 must be **discounted** to express them in terms comparable to dollars in year 1.

As an example, suppose that the interest rate is $i_1 = 5\%$ per year. Assume that a household has \$100 of income in year 1 but plans to spend this income a year later, in year 2. Then the household can buy \$100 of bonds at the start of year 1 and have \$105 available at the start of year 2. Hence, \$100 received in year 1 is worth just as much as \$105 received in year 2. Equivalently, the \$105 from year 2 has to be discounted to correspond to the income needed in year 1 to generate \$105 in year 2. We find this amount by solving the equation:

$$\textit{income needed in year 1 x } (1 + 5\%) = \$105.$$

The required amount of income in year 1 is \$105/1.05 = \$100.

More generally, if we substitute i_1 for 5%, the income for year 2 has to be divided by $1+i_1$ to find the equivalent amount for year 1. Hence, if the wage income received in year 2 is $(w/P)_2 \cdot L$, the present value (or year 1 value) of this income is $(w/P)_2 \cdot L /(1+i_1)$. Economists call the term $1+i_1$ the **discount factor**. When we discount by this factor—that is, divide by $1+i_1$—we determine the present value of year 2's income.

Equation (7.7) shows that we express year 2's wage income as a present value, $(w/P)_2 \cdot L /(1+i_1)$, before combining it with year 1's income, $(w/P)_1 \cdot L$. The sum, $(w/P)_1 \cdot L + (w/P)_2 \cdot L /(1+i_1)$, is the total present value of wage income for years 1 and 2. Similarly, we express year 2's consumption as the present value $C_2/(1+i_1)$ before combining it with year 1's consumption, C_1. The sum, $C_1 + C_2/(1+i_1)$, is the total present value of consumption for years 1 and 2.

2. Choosing consumption: income effects. Thus far, we have considered household budget constraints. Equation (7.7) tells each household the combinations of this year's and next year's consumptions, C_1 and C_2, that satisfy the two-year budget constraint. To understand households' choices of C_1 and C_2, we have to bring in households' **preferences** about consuming at various points in time. To keep things simple, we continue to think about consumption this year and next year, C_1 and C_2.

We assume that households get satisfaction or **utility** from the amounts consumed each year, C_1 and C_2. Other things the same, utility increases if C_1 or C_2 rises. We assume further that households like to consume at similar levels at different points in time, rather than consuming at high levels some of the time and low levels at other times. For example, households prefer having C_1 and C_2 both equal to 100, rather than having C_1 equal to zero and C_2 equal to 200. These preferences give households incentives to **smooth consumption** even when their income is irregular.

Before we work through the formal analysis, we can consider intuitive examples of consumption smoothing. Suppose that a person gets an unexpected windfall of income, perhaps from winning the lottery or getting a surprise check from one's parents. The usual response is to spread the extra money over consumption at various dates, rather than blowing it all on one wild party. Similarly, because people anticipate that their incomes will go down when they retire, they tend to save in advance in order to avoid a sharp drop in consumption during retirement.

Let's return to the two-year budget constraint in equation (7.7) to see how households choose the two consumption levels, C_1 and C_2. The first term on the right-hand side of this budget constraint is the value in year 1 of initial assets, $(1+i_0)\cdot(B_0/P + K_0)$. This term adds to the present value of wage incomes received in years 1 and 2, $(w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1)$. Thus, everything on the right-hand side represents values in year 1—present values—and we can represent this total present value by V:

$$(7.8) \qquad V = (1+i_0)\cdot(B_0/P + K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1).$$

If we substitute this definition of V into equation (7.7), we get

$$(7.9) \qquad C_1 + C_2/(1+i_1) = V - (B_2/P + K_2)/(1+i_1).$$

The last term on the right-hand side is the present value of the real assets held at the end of year 2, $(B_2/P + K_2)/(1+i_1)$. These assets represent a provision for consumption in years 3 and later. We assume, for now, that this term is constant. That is, we analyze the choices of C_1 and C_2 while holding fixed the assets that a household provides for year 3 and beyond.

Suppose that V increases due to a rise in initial assets, $(B_0/P + K_0)$, or wage incomes, $(w/P)_1\cdot L$ *and* $(w/P)_2\cdot L$. Since we are holding fixed the term $(B_2/P + K_2)/(1+i_1)$, equation (7.9) tells us that the total present value of consumption, $C_1 + C_2/(1+i_1)$, rises by the same amount as V. Since households like to consume at similar levels in the two years, we predict that C_1 and C_2 will rise by similar amounts. These responses are called **income effects**. An increase in V, the present value of initial assets and incomes, leads to higher consumption in each year.

3. Choosing consumption: the intertemporal-substitution effect. The income effects that we just studied tell us about the overall level of consumption—for example, the responses of C_1 and C_2 to a change in the present value of initial assets and wage incomes. The other major consideration is how much to consume in one year compared to the other. We have already assumed that households like to have similar levels of C_1 and C_2. However, this preference is not absolute. Households would be willing to deviate from equal consumption levels if there is an economic incentive to deviate. We now show that the interest rate, i_1, provides this incentive.

Consider again the two-year budget constraint:

(7.9) $$C_1 + C_2/(1+i_1) = V - (B_2/P + K_2)/(1+i_1).$$

The left-hand side has the present value of consumption, $C_1 + C_2/(1+i_1)$. Thus, units of C_2 are discounted by $1+i_1$ before adding them to C_1. This discounting means that a unit of C_2 is effectively cheaper than a unit of C_1. The reason is that, if a household defers consumption from year 1 to year 2, it can hold more assets (or borrow less) at the end of year 1. Since each unit of assets becomes $1+i_1$ units in year 2 (see equation [7.5]), one unit less of C_1 can be replaced by $1+i_1$ units more of C_2.

As an example, suppose that you are considering taking a vacation this summer. If the interest rate is $i_1 = 5\%$, you might prefer to postpone the vacation until next summer. The reward is that you could spend 5% more and have a somewhat better vacation. If the interest rate rises to $i_1 = 10\%$, the reward for waiting rises—now you could spend 10% more on the delayed vacation. Thus, our prediction is that the vacation is more likely to be postponed (so that C_1 falls) when the interest rate, i_1, rises.

The general point is that an increase in the interest rate, i_1, lowers the cost of C_2 compared to C_1. That is, a higher i_1 means a greater reward for deferring consumption. Therefore, households respond to an increase in i_1 by lowering C_1 and raising C_2. In other words, a higher interest rate motivates households to substitute consumption over time—away from the present, C_1, and toward the future, C_2.[1] Economists call this response an **intertemporal-substitution effect**.

We can interpret the results in terms of saving by returning to the household budget constraint for year 1:

$$C_1 + (B_1/P + K_1) - (B_0/P + K_0) = (w/P)_1 \cdot L + i_0 \cdot (B_0/P + K_0). \quad (7.2)$$

We know from the intertemporal-substitution effect that an increase in the interest rate, i_1, motivates households to postpone consumption, so that this year's consumption, C_1, falls on the left-hand side. Since the right-hand side of equation (7.2) is fixed, the decline in C_1 must be matched by a rise in this year's real saving, $(B_1/P + K_1) - (B_0/P + K_0)$. That is, the intertemporal-substitution effect motivates households to save more when the interest rate rises. However, our analysis of interest rates is incomplete because we have not considered whether a change in the interest rate has an income effect.

4. The income effect from a change in the interest rate. We can understand the income effect from a change in the interest rate, i_1, by examining the household budget constraint for year 2:

$$C_2 + (B_2/P + K_2) - (B_1/P + K_1) = (w/P)_2 \cdot L + i_1 \cdot (B_1/P + K_1). \quad (7.3)$$

[1] Economists usually assume that households prefer to consume earlier rather than later. In this case, the interest rate, i_1, has to be greater than zero—perhaps 2% per year—to motivate households to choose equal values of C_1 and C_2. If i_1 is greater than 2%, households set C_1 below C_2, whereas if i_1 is less than 2%, households set C_1 above C_2. The main point is still that an increase in i_1 reduces C_1 and raises C_2.

We can see the income effect from i_1 in the term $i_1 \cdot (B_1/P + K_1)$, which gives the income on assets in year 2. We can break this term down into its two parts, $i_1 \cdot (B_1/P)$ and $i_1 K_1$.

Consider first the part $i_1 \cdot (B_1/P)$. This interest income is greater than zero for a lender, which is a household that holds bonds, so that B_1/P is greater than zero. However, this term is less than zero for a borrower, which is a household that issues bonds, so that B_1/P is less than zero. For a lender, the income effect from an increase in i_1 is positive, because the interest income received on a given amount of bonds, B_1/P, is larger. For a borrower, the income effect from an increase in i_1 is negative, because the interest paid on a given amount of debt, B_1/P, is larger. For the economy as a whole, we know that lending and borrowing must balance. In particular, any bond outstanding has both a debtor (the issuer of the bond) and a creditor (the holder of the bond). Therefore, for the average household, B_1/P has to be zero. Hence, for the average household, the income effect from the term $i_1 \cdot (B_1/P)$ is zero.

Households also hold assets in the form of ownership of capital, and the term $i_1 K_1$ in equation (7.3) represents the income received on these assets in year 2. For the economy as a whole, the capital stock, K_1, is, of course, greater than zero. Thus, in contrast to bonds, the average household's holding of claims on capital, K_1, is greater than zero. Therefore, when we consider the term $i_1 K_1$, the income effect from an increase in i_1 is positive.

To put the results together, in the aggregate, the income effect from an increase in i_1 consists of a zero effect from the term $i_1 \cdot (B_1/P)$ and a positive effect from the term $i_1 K_1$. Therefore, the full income effect from an increase in i_1 is positive.

Now we can consider the overall effect from an increase in the interest rate, i_1, on year 1's consumption, C_1. We know that the intertemporal-substitution effect motivates households to reduce C_1. However, an increase in i_1 also has a positive income effect, which motivates households to raise C_1. Therefore, the overall effect from an increase in i_1 on C_1 is ambiguous. Year 1's consumption, C_1, falls if the intertemporal-substitution effect dominates but rises if the income effect dominates. When we apply the theory to particular cases we shall see that we can sometimes resolve the ambiguity by determining whether the income effect is strong or weak.

Empirical evidence on intertemporal substitution of consumption

Our theory predicts that a higher interest rate motivates households to reduce current consumption compared to future consumption. A study by David Runkle (1991) isolated the effect of interest rates on consumption by examining food outlays for 1100 U.S. households from 1973 to 1982. (The data are from the *Panel Study of Income Dynamics* or PSID, conducted at the University of Michigan.) Runkle found that an increase in the annual interest rate by one percentage point raises the typical family's growth rate of consumption by about one-half percentage point per year. The isolation of intertemporal-substitution effects from aggregate consumption data have proven to be difficult, as discussed by Robert Hall (1989). However, a careful statistical study of U.S. non-durable consumption by Joon-Ho Hahm (1998) estimates that a rise in the annual interest rate by one percentage point increases the growth rate of consumption by around one-third percentage point per year.

B. Consumption over many years

Thus far, we considered income and consumption over two years and worked out the two-year budget constraint:

(7.7) $C_1 + C_2/(1+i_1) = (1+i_0)\cdot(B_0/P+K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L /(1+i_1) - (B_2/P+K_2)/(1+i_1)$.

To see how households chose this year's and next year's consumptions, C_1 and C_2, we did not allow the household to change the present value of assets held over to year 3, $(B_2/P + K_2)/(1+i_1)$. These assets are, in fact, not given. A change in $(B_2/P + K_2)$ means that households are providing more or less for consumption in years 3 and beyond. To understand the choice of $(B_2/P + K_2)$, we have to consider consumption and income in future years. The appendix shows in detail how to make this extension. Here we provide an intuitive analysis.

The left-hand side of equation (7.7) is the present value of consumption for years 1 and 2. When we consider many years, the left-hand side becomes the present value of consumption over these many years. The first term added is the present value of year 3's consumption, which is $C_3/[(1+i_1)\cdot(1+i_2)]$. We divide C_3 by $(1+i_1)\cdot(1+i_2)$ because this term measures the cumulation of interest earnings over two years, from year 1 to year 3. That is, one unit of assets in year 1 becomes $1+i_1$ units in year 2, and each of these units becomes $1+i_2$ units in year 3.

If we continue to include years further into the future, we end up with the overall present value of consumption:

$$overall\ present\ value\ of\ consumption = C_1 + C_2/(1+i_1) + C_3/[(1+i_1)\cdot(1+i_2)] + \ldots$$

The ellipses (…) mean that we include the present values of C_4, C_5, and so on. The multi-year budget constraint has this overall present value of consumption on the left-hand side.[2] In contrast, equation (7.7) had this present value only for years 1 and 2.

The right-hand side of equation (7.7) includes the year-one value of initial assets, $(1+i_0)\cdot(B_0/P+K_0)$. This term still appears in the multi-year setting. However, the equation includes the present value of wage incomes only for years 1 and 2. When we consider many years, we end up with the present value of wage incomes over these many years. By analogy to the results for consumption, we end up with

$$overall\ present\ value\ of\ wage\ income = (w/P)_1\cdot L + (w/P)_2\cdot L\ /(1+i_1)$$
$$+ (w/P)_2\cdot L\ /[(1+i_1)\cdot(1+i_2)] + \ldots$$

Again, the ellipses mean that we include the present values of $(w/P)_3\cdot L$, $(w/P)_4\cdot L$, and so on. Notice that the interest-rate terms—needed to calculate present values—are the same as those for consumption.

The final consideration is that the right-hand side of equation (7.7) includes the present value of assets held at the end of year 2, $(B_2/P+K_2)/(1+i_1)$. When we consider many years, this term becomes the present value of assets held in the distant future. Because of the discounting used to calculate present values, we can safely neglect this term. (See the appendix for a discussion.) Therefore, we end up with the following form of the **multi-year budget constraint**:

[2] Year 4's consumption, C_4, is divided by $(1+i_1)\cdot(1+i_2)\cdot(1+i_3)$, and so on.

Key equation:

overall present value of consumption = value of initial assets +

overall present value of wage income

$$C_1 + C_2/(1+i_1) + C_3/[(1+i_1)\cdot(1+i_2)] + \ldots = (1+i_0)\cdot(B_0/P + K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1) + (w/P)_3\cdot L/[(1+i_1)\cdot(1+i_2)] + \ldots \tag{7.10}$$

The multi-year budget constraint allows us to compare the effects of temporary and permanent changes in income. For a temporary change, we can consider an increase in year 1's wage income, $(w/P)_1\cdot L$, by one unit while leaving unchanged the initial assets, $(B_0/P + K_0)$, and the wage incomes for the other years, $(w/P)_2\cdot L$, $(w/P)_3\cdot L$, and so on. One possibility, which satisfies the multi-year budget constraint in equation (7.10), is that a household would spend all of its extra income on year 1's consumption, C_1. However, households tend not to react this way, because they like to have similar levels of consumption in each year. Thus, households typically respond to a rise in $(w/P)_1\cdot L$ by raising consumption by similar amounts in each year, C_1, C_2, C_3, … This response means, however, that consumption in any particular year, such as C_1, cannot increase very much. Therefore, if $(w/P)_1\cdot L$ rises by one unit, C_1 increases by much less than one unit. To put it another way, the **propensity to consume** in year 1 out of an extra unit of year 1's income tends to be small when the extra income is temporary.

We can interpret the results in terms of saving by looking again at the household's budget constraint for year 1:

$$C_1 + (B_1/P + K_1) - (B_0/P + K_0) = (w/P)_1\cdot L + i_0\cdot(B_0/P + K_0). \tag{7.2}$$

We have that $(w/P)_1 \cdot L$ rises by one unit on the right-hand side, whereas C_1 rises by much less than one unit on the left-hand side. Therefore, year 1's real saving, $(B_1/P + K_1) - (B_0/P + K_0)$, rises by nearly one unit on the left-hand side. That is, the **propensity to save** in year 1 out of an extra unit of year 1's income is nearly one when the extra income is temporary. Saving goes up so much because additional assets are needed to provide for the planned increases in consumption in future years.

Consider, as a contrast, a permanent rise in wage income, where $(w/P)_1 \cdot L$, $(w/P)_2 \cdot L$, $(w/P)_3 \cdot L$, … each rise by one unit. The multi-year budget constraint in equation (7.10) shows that it would be possible for households to respond by raising consumption by one unit in each year—in that case, each increase in C_t would match each increase in $(w/P)_t \cdot L$. Moreover, we predict that households would respond roughly this way because this response is consistent with the desire to have similar levels of consumption each year. Thus, the prediction is that the propensity to consume out of an extra unit of year 1's income would be high—close to one—when the extra income is permanent.

For the response of saving, we can again look at equation (7.2). If $(w/P)_1 \cdot L$ rises by one unit on the right-hand side and C_1 rises by roughly one unit on the left-hand side, year 1's real saving, $(B_1/P + K_1) - (B_0/P + K_0)$, changes by little or not at all. In other words, the propensity to save in year 1 out of an extra unit of year 1's income is small when the extra income is permanent. Saving does not change much because, in this case, households do not need additional assets to provide for the planned increases in future consumption. These increases can be paid for by the higher future wage incomes, $(w/P)_2 \cdot L$, $(w/P)_3 \cdot L$, …

Our findings about temporary and permanent changes of income correspond to Milton Friedman's famous concept of **permanent income**.[3] His idea was that consumption depends on a long-term average of incomes—which he called permanent income—rather than current income. If a change in income is temporary, permanent income and, hence, consumption change relatively little. Therefore, as in our analysis, the propensity to consume out of temporary income is small.

Empirical evidence on the propensity to consume

Economists have found strong evidence that the propensity to consume out of permanent changes in income is much larger than that for temporary changes. Some of the clearest evidence comes from special circumstances in which people received windfalls of income, which were clearly temporary and at least partly unanticipated.

One example is the receipt by Israeli citizens of lump-sum, nonrecurring restitution payments from Germany in 1957-58 (see Kreinin [1961] and Landsberger [1970]). The payments were large and roughly equal to an average family's annual income. The typical family increased its consumption expenditure during the year of the windfall by about 20% of the amount received. However, consumption expenditure includes consumer durables,

[3] See Friedman (1957, chapters 2 and 3).

which last for many years and should be regarded partly as saving rather than consumption. Therefore, the true propensity to consume was less than 20%.

Another example is the payment in 1950 to U.S. World War II veterans of an unanticipated, one-time life insurance dividend of about $175, roughly 4% of the average family's annual income. In this case, consumption expenditure rose by about 35% of the windfall (see Bird and Bodkin [1965]). However, since consumption expenditure includes consumer durables, the true propensity to consume was less than 35%.

More general studies of consumer behavior show that the propensity to consume out of permanent changes in income is large and not much different from unity. In contrast, the propensity to consume out of temporary income is about 20 to 30% (see Hall [1989]). Although this response to temporary changes is greater than that predicted by our theory, the important point is that the response of consumption to permanent changes in income is much larger than that to temporary changes.

II. Consumption, Saving, and Investment in Equilibrium

We have discussed how households divide their real income between consumption and real saving. Now we determine the amounts of consumption and real saving that arise when the various markets clear.

Consider again the household budget constraint:

(7.1) $$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L + i\cdot(B/P + K).$$

If we separate real asset income, $i\cdot(B/P + K)$, into its two parts, $i\cdot(B/P)$ and iK, we get the revised budget constraint:

$$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L + i\cdot(B/P) + iK.$$

We know from chapter 6 that the interest rate, i, equals the rate of return on capital:

(6.4) $$i = (R/P - \delta).$$

If we substitute $(R/P - \delta)$ for i in the iK term of the budget constraint, we have

$$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L + i\cdot(B/P) + (R/P)\cdot K - \delta K.$$

If we apply this condition to the aggregate of households, we know that the total quantity of bonds, B, must be zero. That is, when the bond market clears, households in the aggregate hold a zero net quantity of bonds. This condition also implies that the change in aggregate bond holdings, ΔB, equals zero. If we substitute $B = 0$ and $\Delta B = 0$ into the equation, we get that, in the aggregate, the household budget constraint becomes

$$C + \Delta K = (w/P)\cdot L + (R/P)\cdot K - \delta K.$$

We know from chapter 6 that, when the labor and rental markets clear, the total payments to factors, $(w/P)\cdot L$ for labor plus $(R/P)\cdot K$ for capital, equals real GDP, $Y = A\cdot F(K,L)$. (See the appendix to chapter 6.) If we substitute $A\cdot F(K,L)$ for $(w/P)\cdot L + (R/P)\cdot K$ in the equation, we find that the aggregate household budget constraint is

(7.11) $$C + \Delta K = A\cdot F(K,L) - \delta K$$

consumption + net investment = real GDP – depreciation

= real net domestic product.

In equation (7.11), the real net domestic product, $A \cdot F(K,L) - \delta K$, is determined, for a given technology level, A, by the given values of K and L. The economy's net investment, ΔK, is therefore determined by households' choices of consumption, C. Given the real net domestic product, one unit more of consumption, C, means one unit less of net investment, ΔK. In the next section, we investigate how much households choose to consume, C, given that the interest rate, i, is determined from equation (6.4) to equal the rate of return on capital, $MPK - \delta$. This choice of C then determines ΔK from equation (7.11).

Notice that, when the bond market clears, net investment, ΔK, equals economy-wide real saving. For an individual household, real saving equals $(1/P) \cdot \Delta B + \Delta K$—the change in the real value of assets held as bonds or capital. However, for the economy as a whole, ΔB equals zero, and real saving equals ΔK.

III. Economic Fluctuations

Now we want to use the model to understand the economy's short-term economic fluctuations. These fluctuations are often called **business cycles**. We think of movements in real GDP as the key indicator of whether an economy is expanding or contracting. Thus, to understand the nature of economic fluctuations, we look first at real GDP for the United States in the post-World War II period.

A. Cyclical behavior of real GDP—recessions and booms

The blue graph in Figure 7.2 shows the logarithm of U.S. real GDP on a quarterly, seasonally-adjusted basis from 1947.1 to 2004.2. To visualize the fluctuations in real GDP, we filter out the trend. The trend, shown in red, is a smooth curve drawn through the real GDP numbers.[4]

Figure 7.2 makes clear that the most important property of U.S. real GDP from 1947 to 2004 is the overall upward movement or trend, shown by the red curve. This trend determines how the typical person's standard of living in 2004 compares with the standard 10, 20, or 50 years earlier. The trend reflects the economy's growth over the long run. In chapters 3-5, we analyzed the sources of this economic growth.

Now we are focusing on the fluctuations of real GDP around its trend. In Figure 7.2, we represent these fluctuations by the deviations of the log of real GDP (the blue graph) from its trend (the red curve). Although these fluctuations are small, relative to the trend movements, the fluctuations also influence the typical person's well-being, especially in the short run. Moreover, there is no question that most economic discussions in the news media focus on fluctuations rather than trends. Probably this is because trends represent long-term forces that are usually not news, whereas fluctuations reflect recent events that do constitute news.

If we calculate the difference between the log of real GDP and its trend from Figure 7.2, we get the variable graphed in Figure 7.3. We call this variable the **cyclical part of real GDP**. The variability of this cyclical part is a good way to gauge the extent of economic fluctuations. To get a quantitative measure, we use the standard deviation.

---

[4] The trend line shown in the figure is called a Hodrick-Prescott filter (H-P filter), named after the economists Robert Hodrick and Edward Prescott. The general idea is to determine the position of the trend to fit the movements in real GDP without fluctuating "too much." The procedure allows the slope of the trend line to change over time in response to changes in the long-run growth rate of real GDP.

For the variable shown in Figure 7.3 from 1947.1 to 2004.3, this value is 1.7%. Hence, the typical range of fluctuations of U.S. real GDP in the post-World War II period is between 1.7% below and 1.7% above its trend.[5]

The low points or troughs in Figure 7.3 pick out the nine U.S. **recessions** from 1947 to 2004. These recessions occurred in 1949, 1954, 1958, 1961, 1970, 1975, 1982-83, 1990-91, and 2001-02. The departures of real GDP from trend ranged from the mild recessions of 2001-02 (maximum of 1.6% shortfall from trend), 1990-91 (1.6%), and 1961 (2.7%) to the more severe recessions of 1949 (6.2%), 1982-83 (4.7%), 1958 (4.2%), and 1975 (3.8%). The recession dates shown in the graph correspond well to those designated by the semi-official arbiter of recessions, the National Bureau of Economic Research or NBER.[6] We can also see the boom periods from the high points or peaks. The most recent peak was at 2.3% above trend in the second quarter of 2000—this peak coincided with large investments in the Internet and other forms of advanced technology.

Recessions in the long-term U.S. history

The recessions of the post-World War II period, shown in Figure 7.3, were not nearly as large as some earlier ones, notably the Great Depression in the early 1930s. We can measure earlier recessions using the reasonably good annual data

[5] If a variable is normally distributed (which is a reasonable approximation for the log of real GDP), about two-thirds of the time the variable is between one standard deviation below and one standard deviation above its mean. About 95% of the time, the variable is between two standard deviations below and two standard deviations above its mean.
[6] The main difference is that the NBER designated the brief dip in real GDP in 1980 as another recession. The shortfall of real GDP from trend in 1980.3 reached 1.2%, about the same as in 1996.1, which the NBER did not call a recession.

on U.S. real GDP available since 1869. (The data before 1929 are for gross national product or GNP.) Figure 7.4 shows recessions for the period 1869 to 2003. The method is same as the one used for post-World War II data in Figure 7.3, except that the data are annual, rather than quarterly. The years marked off in Figure 7.4 are those for which real GDP (or GNP) fell short of its trend by at least 3%. With this cutoff, some of the post-World War II recessions—including those for 1990-91 and 2001-02—are too mild to be marked.

The Great Depression dwarfs any of the recessions experienced since World War II. In the worst year, 1933, real GDP fell short of trend by 19%. Two other large recessions during the inter-war period were those of 1920-22 and 1938-40—real GDP was 9% below trend in 1921 and again in 1938.

The years before World War I do not contain any recessions as large as those of the inter-war years. In fact, the period from 1869 to 1914 does not differ greatly from that since 1947 in terms of the extent of economic fluctuations. It is hard to be sure because the national-accounts data are less reliable for the long ago past.[7]

[7] Christina Romer's (1986, 1987) detailed analysis suggests that the pre-World War I period was only slightly more variable than the post-World War II period in terms of the extent of economic fluctuations.

B. Real business cycle theory

To model economic fluctuations, we start with the assumption that fluctuations occur because of shocks that hit the economy. Further, we assume that we can explain the fluctuations as short-term responses to these shocks. The main thing that makes the analysis short term is that we assume, as an approximation, that we can hold fixed the stock of capital, K. That is, in thinking about the relatively brief duration of a recession or a boom, we do not allow enough time to elapse for the changes in machines and buildings—the goods included in the stock of capital, K—to be significant. In contrast, for long-term analyses of economic growth, as in chapters 3-5, the changes in the stock of capital were a major part of the story.

In one important theory of economic fluctuations, called **real business cycle theory**, the shocks to the economy are assumed to be real (as opposed to monetary, which we consider in chapters 15 and 16). The form of real shock that is most commonly assumed is a change in the technology level, A. However, for this approach to have a chance to work, we have to allow for shocks to A to be sometimes positive—which will generate booms—and sometimes negative—which will generate recessions.

If we think of A as the technology level, it is easy to imagine positive shocks, for example, from discoveries of new goods or methods of production. Examples mentioned in our study of technological progress in chapter 5 are the invention and adaptation of electric power, the transistor, computers, and the Internet. Significant innovations in production and management include the assembly line, just-in-time inventory control, and Wal-Mart's model of business management.[8] However, many smaller inventions and innovations contributed to the economic booms shown in Figure 7.3.

If we view *A* as literally the technology level, it is hard to imagine important negative shocks, because producers normally would not forget previous technological advances. However, we mentioned in our study of the Solow growth model in chapter 5 that changes other than technological shifts can affect productivity and, thereby, influence the economy in ways similar to changes in the technology level. These other kinds of changes include shifts in legal and political systems, changes in the degree of competition, and variations in the volume of international trade. One interesting case is the rise of the Organization of Petroleum Exporting Countries (OPEC) in 1973, which changed the competitive structure of the international petroleum market and led, for awhile, to large increases in the price of oil. Other adverse events that have effects similar to reductions in *A* are harvest failures, wartime destruction, natural disasters, and strikes.

In our analysis of economic fluctuations, we shall adopt a broad view of *A* that encompasses these various examples. In this case, shocks to *A* can sometimes be positive and sometimes negative. However, we shall find it convenient still to refer to *A* as the "technology level." We should remember that our concept of *A* includes an array of economic and political variables that affect productivity.

Now we consider what our model says about the short-run effects from a shift in the technology level, *A*. Recall that real GDP, *Y*, is given by the production function,

$$Y = A \cdot F(K,L). \tag{3.1}$$

[8] For a more complete list of examples, see the discussion of innovations in Jones (2003).

We mentioned that we will treat the capital stock, K, as fixed in the short run. For now, we continue to assume that labor, L, is also fixed. In this case, changes in Y will reflect only changes in A. In chapter 8, we make an important extension to allow for short-run variations in labor and capital services.

In practice, many shifts to A will not be observable, that is, we will not be able to tell what changes have occurred in the variables that influence the economy's productivity. The problem is that if we are free to make assumptions about which unobservable changes have occurred, we will be able to match any observed fluctuations in real GDP, such as those shown in Figure 7.3. Given how easy it is to fit these data, we should not give our model any credit for "explaining" the fluctuations in real GDP in this way. The real challenge for the model is to predict how other macroeconomic variables move along with real GDP during economic fluctuations. As an example, we want to see what the model predicts for the changes of consumption and investment during booms and recessions. Similarly, we can assess the behavior of the real wage rate, the real rental price of capital, and the interest rate. We now start our consideration of these macroeconomic variables.

1. The marginal product of labor and the real wage rate. We know from the production function in equation (3.1) that an increase in the technology level, A, raises the marginal product of labor, MPL, for given K and L. We show the effects from a higher schedule for the MPL in Figure 7.5. We consider two technology levels, A_1 and A_2, where A_2 is greater than A_1. We assume throughout that the capital stock is fixed at K. The downward-sloping blue curve shows how the MPL varies with L when the technology level is A_1. If the real wage rate is w/P, on the vertical axis, the quantity of labor demanded is the amount L_1^d on the horizontal axis. The downward-sloping red curve shows the MPL when the technology level is A_2. The MPL is higher at any given L on this second curve than on the first one. Therefore, at the given real wage rate, w/P, the quantity of labor demanded L_2^d, on the horizontal axis, is greater than L_1^d.

In Figure 7.6, we assume that labor supply is fixed in the short run at L, shown on the horizontal axis. If the technology level is A_1, so that the MPL is given by the downward-sloping blue curve, the market-clearing real wage rate is $(w/P)_1$, shown on the vertical axis. The value $(w/P)_1$ equals the MPL evaluated at L, when the technology level is A_1 (and the capital stock is fixed at K). In contrast, if the technology level is A_2, the MPL is given by the downward-sloping red curve. In this case, the market-clearing real wage rate equals $(w/P)_2$, shown on the vertical axis. Since the MPL is higher, at the given L, on the red curve than on the blue one, the market-clearing real wage rate is higher. That is, $(w/P)_2$ is greater than $(w/P)_1$.

One way to think about the result is that, at the initial real wage rate, $(w/P)_1$, the rise in the MPL means that the quantity of labor demanded, L^d, exceeds the quantity supplied, which is fixed at L. Therefore, employers (households in their role as business managers) compete for the scarce labor and thereby drive up the real wage rate to $(w/P)_2$.

We conclude that an increase in the technology level, A, raises the real wage rate, w/P. Hence, the model predicts that an economic boom—where real GDP is high because A is high—will have a relatively high w/P. In contrast, a recession will have a relatively low w/P.

2. The marginal product of capital, the real rental price, and the interest rate. We know from the production function in equation (3.1) that an increase in the technology level, A, also raises the marginal product of capital, MPK, for given K and L. We show the effects from a higher MPK in Figure 7.7. This figure again considers two technology levels, A_1 and A_2, where A_2 is greater than A_1. We assume that labor input is fixed at L. The downward-sloping blue curve shows how the MPK varies with K when the technology level is A_1. If the real rental price is R/P, shown on the vertical axis, the quantity of capital demanded is the amount K_1^d on the horizontal axis. The downward-sloping red curve shows the MPK when the technology level is A_2. The MPK is higher at any given K on this second curve than on the first one. Therefore, at the given real rental price, R/P, the quantity of capital demanded K_2^d, on the horizontal axis, is greater than K_1^d.

In Figure 7.7, we assume that the supply of capital services is fixed in the short run at K, shown on the horizontal axis. If the technology level is A_1, so that the MPK is given by the downward-sloping blue curve, the market-clearing real rental price is $(R/P)_1$, shown on the vertical axis. The value $(R/P)_1$ equals the MPK evaluated at K, when the technology level is A_1 (and labor input is fixed at L). In contrast, if the technology level is A_2, the MPK is given by the downward-sloping red curve. In this case, the market-clearing real rental price equals $(R/P)_2$, shown on the vertical axis. Since the MPK is higher, at the given K, on the red curve than on the blue one, the market-clearing real rental price is higher. That is, $(R/P)_2$ is greater than $(R/P)_1$.

We conclude that an increase in the technology level, A, raises the real rental price of capital, R/P. Hence, the model predicts that an economic boom—where real GDP is high because A is high—will have a relatively high R/P. In contrast, a recession will have a relatively low R/P.

Recall that the interest rate is given by

(6.4) $$i = R/P - \delta$$

$$= MPK\ (evaluated\ at\ K) - \delta.$$

We know that an increase in the technology level raises the marginal product of capital, MPK, at a given K, and thereby increases the real rental price of capital, R/P. Therefore, we also have that the interest rate, i, rises. Hence, the model predicts that an economic boom will have a relatively high interest rate, whereas a recession will have a relatively low interest rate.

3. Consumption, saving, and investment. Now we consider how households choose consumption. We know that a rise in the technology level, A, raises the interest rate, i. We also know that a higher i motivates households to defer consumption from the present to the future (the intertemporal-substitution effect). On this ground, we predict that the rise in A would reduce current consumption. However, our analysis of consumption is incomplete, because we have to allow for income effects.

Consider again the household budget constraint:

(7.1) $$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L + i\cdot(B/P + K).$$

Income effects enter through real wage income, $(w/P)\cdot L$, and real asset income, $i\cdot(B/P + K)$. We know that an increase in A raises real wage income, because w/P rises and L does not change. We also know that an increase in A raises real asset income, because i rises, B/P is unchanged (at zero in the aggregate), and K does not change in the short run. Therefore, an increase in A raises overall household real income.

Another way to see the effect of an increase in A on overall household income is to use the aggregate budget constraint that applies when the markets for bonds, labor, and capital services clear:

$$(7.11) \qquad C + \Delta K = A \cdot F(K,L) - \delta K.$$

Since depreciation, δK, is fixed in the short run, the income effect from a change in A boils down to its effect on real GDP, $A \cdot F(K,L)$. Since an increase in A raises real GDP for given K and L, we again see that a rise in A raises overall real income.

The increase in real income motivates households to raise current consumption (as well as consumption in future periods). This response is the familiar income effect. This effect works against the intertemporal-substitution effect, which tended to reduce current consumption. Therefore, we are unsure whether an increase in the technology level, A, leads, on net, to more or less current consumption, C. The net change in C depends on whether the income effect is stronger or weaker than the intertemporal-substitution effect.

We can say more because the size of the income effect depends on how long the change in the technology level, A, lasts. For the rest of this section, we assume that the change in A is permanent. This situation would apply for a literal technological improvement—because producers tend not to forget technological advances. In this case, the increases in real wage and asset incomes tend also to be permanent. Therefore, if we return to our discussion of the propensity to consume from Section I.B, we should consider the case in which real wage incomes—$(w/P)_1 \cdot L$, $(w/P)_2 \cdot L$, …—rise by similar amounts each year. (In the present case, real asset incomes will also rise each year.) Therefore, we predict that the propensity to consume will be close to one. That is, if an increase in A raises real GDP, $A \cdot F(K,L)$, by one unit, then—from the standpoint of the income effect—current consumption, C, would rise by roughly one unit.

To get the overall effect on current consumption, we have to balance the income effect—whereby consumption rises by roughly as much as real GDP—against the intertemporal-substitution effect, which lowers current consumption. Quantitative estimates of the intertemporal-substitution effect show that it is weaker in magnitude than this large income effect. Hence, when the increase in A is permanent, we predict that current consumption will rise.

We also know that the increase in current consumption will be less than the increase in real GDP. The reason is that the income effect raises consumption by roughly as much as real GDP, but the intertemporal-substitution effect lowers current consumption. Therefore, as long as the intertemporal-substitution effect operates at all, the increase in current consumption will be less than the increase in real GDP.

If we look again at equation (7.11), we have found that current consumption, C, rises but by less than the increase in real GDP, $A \cdot F(K,L)$. Therefore, net investment, ΔK, must increase. In other words, the increase in real GDP shows up partly as more consumption, C, and partly as more net investment, ΔK. Recall that ΔK equals real saving for the whole economy. Therefore, we also have that a permanent improvement in the technology level, A, leads to an increase in real saving.

C. Matching the theory with the facts

We have seen that our model makes a number of predictions about how fluctuations in macroeconomic variables match up with fluctuations in real GDP. We evaluate here the predictions for consumption, investment, the real wage rate, the real rental price of capital, and the interest rate. We focus our attention on U.S. data since 1954. From the perspective of Figures 7.2 and 7.3, we are leaving out the period from 1947 to 1953. That period is unusual because it is heavily influenced by the aftermath of World War II and the Korean War. We consider the economic effects of wartime in chapter 12.

1. Consumption and investment. We can measure consumption, *C*, from the national-income accounts by real consumer expenditure. The nominal amount of this expenditure accounted, on average, for 64% of nominal GDP in the United States from 1954.1 to 2004.3. We can calculate the cyclical part of the logarithm of real consumer expenditure by using the method applied to real GDP in Figure 7.2. The result is the blue graph in Figure 7.9. We also show as the red graph the cyclical part of the log of real GDP (from Figure 7.3). Two important findings emerge. First, real consumer expenditure typically fluctuates in the same direction as real GDP.[9] When a variable fluctuates, like real consumer expenditure, in the same direction as real GDP we say that the variable is **procyclical**. Second, the extent of the fluctuations in real consumer expenditure is proportionately smaller than that in real GDP. From 1954.1 to 2004.3, the standard deviation of the cyclical part of real consumer expenditure is 1.2%, compared with 1.6% for the cyclical part of real GDP.

[9] From 1954.1 to 2004.3, the correlation of the cyclical part of real consumer expenditure with the cyclical part of real GDP is 0.87.

We can measure gross investment, *I*, from the national income accounts by real gross domestic private investment. The nominal amount of this expenditure accounted, on average, for 16% of nominal GDP from 1954.1 to 2004.3. We can again use the method from Figure 7.2 to calculate the cyclical part of the log of real gross investment. The result is the blue graph in Figure 7.10. The cyclical part of the log of real GDP is again the red graph. One finding is that, like real consumer expenditure, real gross investment is procyclical, that is, it typically fluctuates in the same direction as real GDP.[10] Another finding is that real gross investment is proportionately far more variable than real GDP. In terms of the standard deviations of the cyclical parts, the one for gross investment is 7.3%, compared to only 1.6% for GDP and 1.2% for consumer expenditure.[11] The extreme volatility of investment means that it represents far more of the cyclical changes in real GDP than we would expect from the average ratio of gross investment to GDP (16%).

Going back to the model, we find that permanent shifts in the technology level, *A*, match up well with some of the patterns described in Figures 7.8 and 7.9. Specifically, increases in *A* generate economic booms, where real GDP increases, and these increases show up partly as more consumption and partly as more investment. In reverse,

[10] From 1954.1 to 2004.3, the correlation of the cyclical part of real gross domestic private investment with the cyclical part of real GDP is 0.92.

[11] Consumer expenditure includes purchases of consumer durables, such as automobiles, furniture, and appliances, as well as spending on non-durables and services. We should think of consumer durables as forms of capital owned by households. Therefore, we should view the purchases of these durables as forms of gross investment. Hence, we could combine the purchases of consumer durables with gross investment to get a broader measure of investment. We would then represent consumption by a narrower measure—real consumer expenditure on non-durables and services. If we make these changes, we find that the cyclical part of the narrower measure of real consumer expenditure is proportionately less variable than the one shown in Figure 7.9. That is, we get a stronger pattern of consumption being less variable than real GDP.

decreases in A create recessions, where real GDP, consumption, and investment all decline.

Consider now whether the model explains why investment is proportionately far more variable than consumption. Recall that, because the changes in the technology level, A, are permanent, the income effects are strong. On this ground, consumption would change by roughly the same amount as real GDP. However, we also found that an increase in A led to a rise in the interest rate, which reduced current consumption and raised current real saving. This effect meant that, during a boom, consumption would rise by less than real GDP. Analogously, during a recession, consumption would fall by less than real GDP. Thus, to match the finding that consumption is less variable than real GDP, the model relies on the intertemporal-substitution effect from the interest rate. One problem, however, is that empirical studies have not found strong evidence for a substantial intertemporal-substitution effect on consumption and saving. Therefore, it may be important to find other reasons why consumption is less variable than real GDP. We explore an important reason in section D.

2. The real wage rate. The model predicts that the real wage rate, w/P, will be relatively high in booms and relatively low in recessions. Our best measure of the nominal wage rate, w, is the average hourly nominal earnings of production workers in the total private economy. We can measure the real wage rate, w/P, by dividing the earnings measure by the deflator for the gross domestic product, which is a broad measure of the price level. We can detrend the logarithm of this real wage variable by the procedure used for real GDP in Figure 7.2. The result is the cyclical part of the real wage rate shown as the blue graph in Figure 7.11. We again show as the red graph the cyclical part of the log of real GDP (from Figure 7.3). We see that the real wage rate is procyclical—it tends to be above trend during booms and below trend during recessions.[12] This finding supports the model, which predicted that the real wage rate would be relatively high in booms and relatively low in recessions.

3. The real rental price. The model predicts that the real rental price of capital, *R/P*, will be relatively high in booms and relatively low in recessions. The main problem in testing this proposition is that the rental price is difficult to measure for the whole economy. The reason is that most forms of capital—such as structures and equipment owned by corporations—are not explicitly rented out. These types of capital are typically used by their owners. In effect, businesses rent capital to themselves, but we cannot observe the implicit rental price, *R/P*, associated with this capital.[13]

Casey Mulligan, in unpublished research, has estimated the implicit real rental price, *R/P*, applicable to capital owned by the corporate sector of the U.S. economy. He made this calculation by dividing an estimate of the total payments to corporate capital by an estimate of the total quantity of this capital.[14] Figure 7.12 shows as the blue graph the cyclical part of *R/P*. We again show as the red graph the cyclical part of the log of real

[12] From 1954.1 to 2004.3, the correlation of the cyclical part of the real wage rate with the cyclical part of real GDP is 0.64. The real wage rate is, however, proportionately much less variable than real GDP. The standard deviation of the cyclical part of the real wage rate is 0.7%, compared with 1.6% for real GDP. The results are similar if we calculate the real wage rate by dividing nominal earnings by the consumer price index (CPI), instead of the GDP deflator.
[13] The national income accounts do include an implicit rental income on owner-occupied housing.
[14] For a discussion of the methodology, see Mulligan (2002). His real rental price is calculated net of corporate taxes.

GDP. We see that *R/P* is procyclical.[15] This finding supports the model, which predicted that *R/P* would be relatively high in booms and relatively low in recessions.

4. The interest rate. The model also predicts that booms, when *A* increases, will have high interest rates, whereas recessions, when *A* decreases, will have low interest rates. This pattern looks right—that is, interest rates tend to be above trend during booms and below trend during recessions.[16] However, to get a reliable picture, we have to consider additional factors, especially monetary policy and inflation, that have substantial influences on interest rates. We consider inflation in chapter 11.

D. Temporary changes in the technology level

In our model, we assumed that all changes in the technology level, *A*, were permanent. This assumption is reasonable for literal discoveries of technology but would be less appropriate for some other interpretations of *A*. For example, if the changes in *A* represent harvest failures or strikes, they would be temporary. To consider these cases, we now redo the analysis for a case in which the change in *A* is temporary. To be concrete, we think of the change in *A* as lasting for one year.

If *A* increases temporarily, real GDP, $A \cdot F(K,L)$, still rises for given values of *K* and *L*. The marginal product of capital, MPK, and the interest rate, *i*, also rise as before. The intertemporal-substitution effect from the higher *i* still motivates households to reduce current consumption, *C*, and raise current real saving.

[15] From 1954.1 to 2003.4 (over which the data on the real rental price are available), the correlation of the cyclical part of the real rental price with the cyclical part of real GDP is 0.52.
[16] From 1954.1 to 2004.3, the correlation of the cyclical part of the interest rate on 3-month U.S. Treasury Bills with the cyclical part of real GDP is 0.38.

The net change in current consumption, C, again depends on the income effect. In our previous case—where the increase in the technology level, A, was permanent—we argued that the income effect tended to raise consumption by about as much as the increase in real GDP. This change worked against the intertemporal-substitution effect, which tended to reduce current consumption. In the case of a permanent rise in A, we argued that the income effect was strong enough to more than offset the intertemporal-substitution effect. Therefore, current consumption increased overall.

The new consideration is that the increase in the technology level, A, is temporary and, hence, the income effect is weak. Thus, the income effect now raises current consumption, C, by only a small amount. The intertemporal-substitution effect tends again to reduce current consumption. Since the income effect is weak, it is now uncertain whether the income effect is stronger or weaker than the intertemporal-substitution effect. Hence, current consumption may rise or fall. In any event, current consumption surely rises by far less than real GDP.

Consider again the aggregate budget constraint that applies when the markets for bonds, labor, and capital services clear:

$$C + \Delta K = A \cdot F(K,L) - \delta K. \tag{7.11}$$

We now have that real GDP, $A \cdot F(K,L)$, rises, but consumption, C, either falls or rises by a small amount. Hence, net investment, ΔK, rises by nearly as much as—or possibly by even more than—real GDP. The model therefore predicts that an economic boom would feature high real GDP and investment. However, consumption would rise by at most a small amount. Conversely, a recession would have low real GDP and investment, but consumption would decline by at most a small amount.

These patterns conflict with observations, because consumption is clearly procyclical—it rises above trend during booms and falls below trend in recessions. Moreover, although consumption is proportionately less variable than real GDP, the changes in consumption during booms and recessions are substantial. Thus, if the underlying shocks were purely temporary changes in the technology level, *A*, the model would not explain the observed fluctuations in consumption. Our conclusion is that we cannot rely on temporary changes in the technology level, *A*, as the main source of economic fluctuations. However, the model does work better if we assume that the changes in *A* are less than fully permanent, even if not purely temporary.

Consider again the finding that consumption is proportionately less variable than real GDP. When we assumed that the changes in the technology level, *A*, were permanent, the model could explain the smaller variability of consumption only if the intertemporal-substitution effect on consumption and saving were substantial. However, if the changes in *A* are less than fully permanent, we have another reason for why consumption would be less variable than real GDP. If a change in *A* lasts for a long time, but not forever, the income effect will be strong. However, the income effect will not be strong enough to raise consumption by as much as the change in real GDP. Thus, consumption may be proportionately less variable than real GDP even if the intertemporal-substitution effect on consumption and saving does not operate. This reasoning suggests that the model works best to fit the facts when the underlying shocks to *A* are long-lasting but less than fully permanent. This form of shock to technology has, in fact, typically been assumed in models of real business cycles.

III. Summing Up

We have shown that the model—a version of a real business-cycle model—can go a long way in matching observed economic fluctuations. The key assumption is that economic fluctuations originate from long-lived, but less than permanent, shifts to the technology level, *A*. More generally, we can interpret shifts in *A* as a variety of real disturbances that influence an economy's productivity. With these types of shocks, the model explains why consumption, investment, the real wage rate, and the real rental price of capital are all procyclical. The model can also explain why investment is proportionately far more variable than consumption.

The most important feature of economic fluctuations that we have neglected is the tendency for labor input, *L*, to be procyclical. Specifically, employment and worker-hours tend to be high in booms and low in recessions. The model could not explain this pattern because we assumed that the quantity of labor supplied, L^s, was fixed. In the next chapter, we allow for variable labor supply and then examine whether the model can explain the observed fluctuations in labor input.

Questions and Problems

Mainly for review

7.1. Derive the two-year household budget constraint shown in equation (7.7). According to this constraint, if a household reduces this year's consumption, C_1, by one unit, how much would next year's consumption, C_2, rise (if nothing else changes in the equation)?

7.2. Show how taking a present value gives different weights to income and consumption in different years. Why is a unit of real income in the present more valuable than a unit of real income next year? Why is a unit of consumption next year cheaper than a unit this year?

7.3. What factors determine whether the propensity to consume out of an additional unit of income is less than or equal to one? Can the propensity be greater than one?

7.4. Discuss the effects on this year's consumption, C_1, from the following changes:

a. An increase in the interest rate, i_1.

b. A permanent increase in real wage income, $(w/P)\cdot L$.

c. An increase in current real wage income, $(w/P)_1\cdot L$, but no change in future real wage incomes.

d. A one-time windfall, which raises initial assets, $(B_0/P + K_0)$.

Problems for discussion

7.x Permanent income

The idea of permanent income is that consumption depends on a long-run average of income, rather than current income. Operationally, we can define permanent income to be the hypothetical, constant income that has the same present value as a household's sources of funds on the right-hand side of the multi-year budget constraint in equation (7.10).

a. Use equation (7.10) to get a formula for permanent income, when evaluated in year 1.

b. What is the propensity to consume out of permanent income?

c. If consumption, C_t, is constant over time, what is the value of permanent income?

7.x. A change in population

Assume a one-time decrease in population, possibly caused by an onset of plague or a sudden out-migration.

a. Use a variant of Figure 7.6 to determine the effects on the labor market. What happens to labor input, L, and the real wage rate, w/P?

b. Use a variant of Figure 7.8 to determine the effects on the market for capital services. What happens to capital input, K, and the real rental price, R/P? What happens to the interest rate, i?

c. What happens to output, Y, and consumption, C? What happens to investment, I? What happens over time to the stock of capital, K?

7.x. A change in the capital stock

Assume a one-time decrease in the capital stock, possibly caused by a natural disaster or an act of war. Assume that population does not change.

a. Use a variant of Figure 7.8 to determine the effects on the market for capital services. What happens to capital input, K, and the real rental price, R/P? What happens to the interest rate, i?

b. Use a variant of Figure 7.6 to determine the effects on the labor market. What happens to labor input, L, and the real wage rate, w/P?

c. What happens to output, Y, and consumption, C? What happens to investment, I? What happens over time to the stock of capital, K?

7.x. A shift in desired saving

Suppose that households change their preferences so that they wish to consume more and save less in the current year. That is, current consumption, C_1, rises for a given interest rate and for given current and future income.

a. What happens to labor input, L, capital input, K, and output, Y?

b. What happens to the real wage rate, w/P, the real rental price, R/P, and the interest rate, i?

c. What happens to consumption, C, and investment, I? What happens over time to the stock of capital, K?

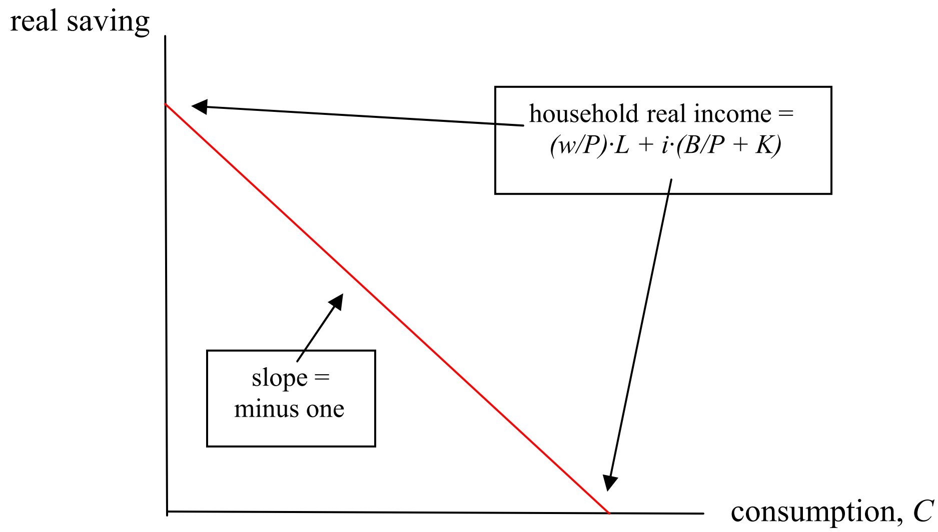
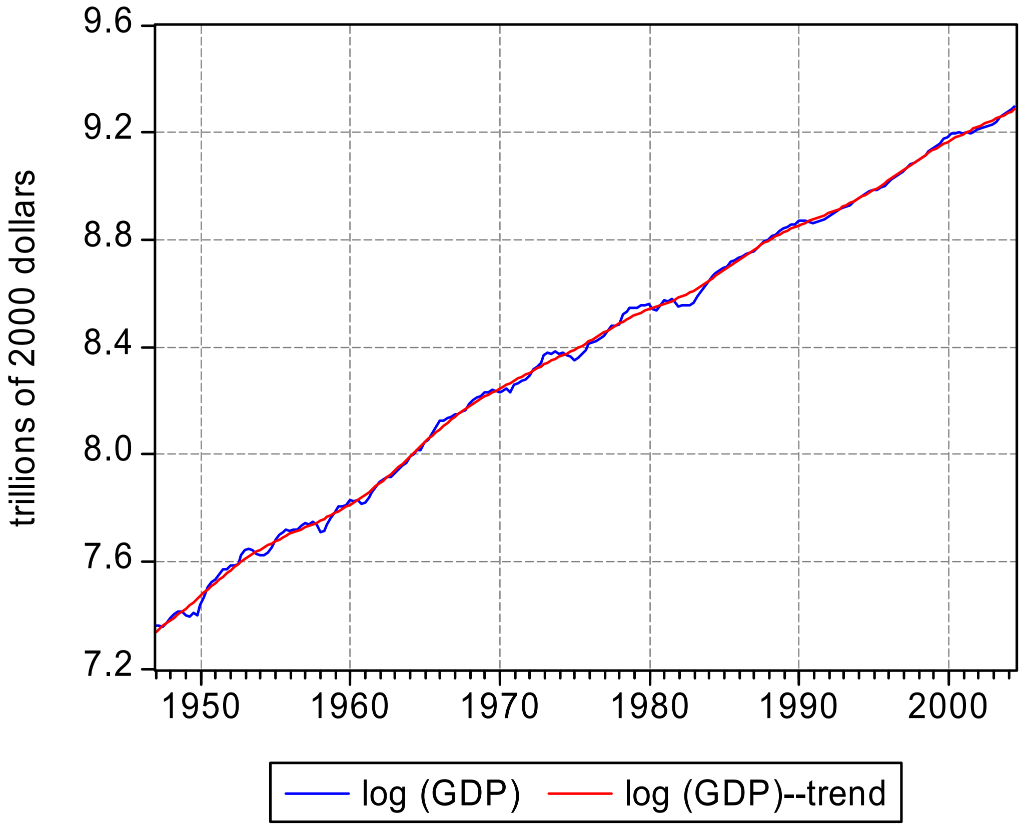


Figure 7.1

The Household Budget Constraint

Households have a given total of real income, $(w/P)\cdot L + i\cdot(B/P + K)$. This total must be divided between consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$. Thus, if real saving is zero, C equals the total of real income along the horizontal axis. If C is zero, real saving equals the total of real income along the vertical axis. The budget constraint in equation (7.1) allows the household to select any combination of consumption and real saving along the red line, which has a slope of minus one. Along this budget line, one unit less of consumption corresponds to one unit more of real saving.



.

Figure 7.2

Trend of U.S. Real GDP, 1947-2004

The blue graph shows the logarithm of U.S. real GDP in trillions of 2000 dollars. The data are quarterly, seasonally-adjusted values from 1947.1 to 2004.2. The red curve is a smooth trend drawn through the GDP data.

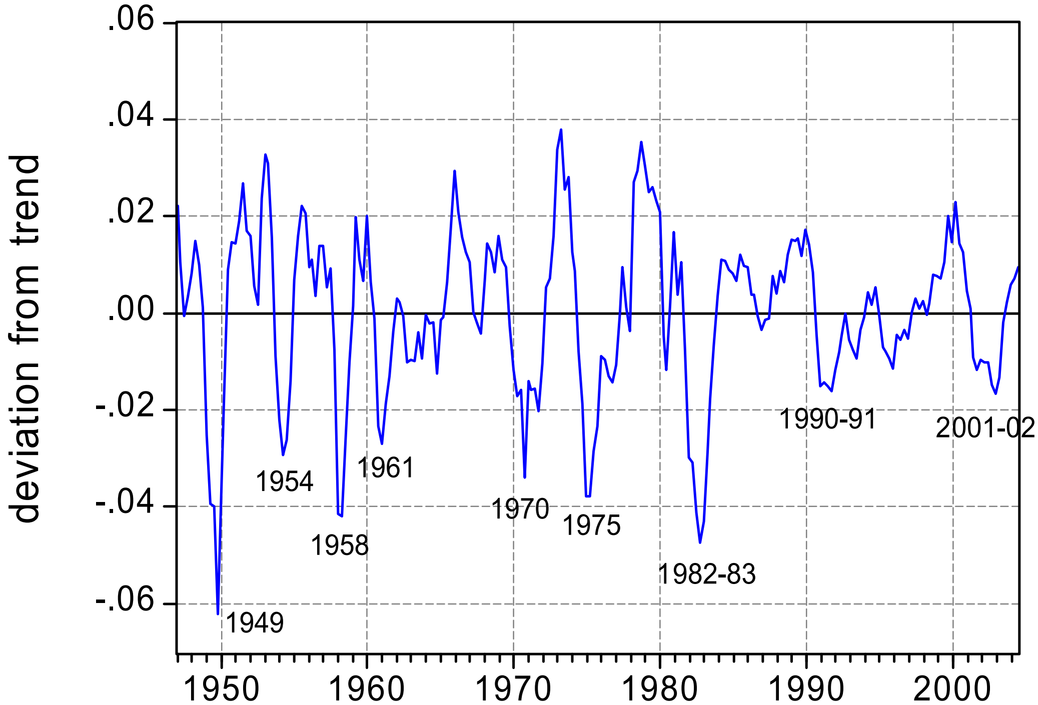


Figure 7.3

Cyclical Part of U.S. Real GDP. 1947-2004

The graph plots the difference between the log of real GDP (the blue graph in Figure 7.2) and its trend (the red graph in Figure 7.2). The resulting series—the cyclical part of real GDP—shows the deviations of the log of real GDP from its trend. As an example, 0.02 means that real GDP is 2% above its trend. The low points pick out the recession years, which are labeled in the graph.

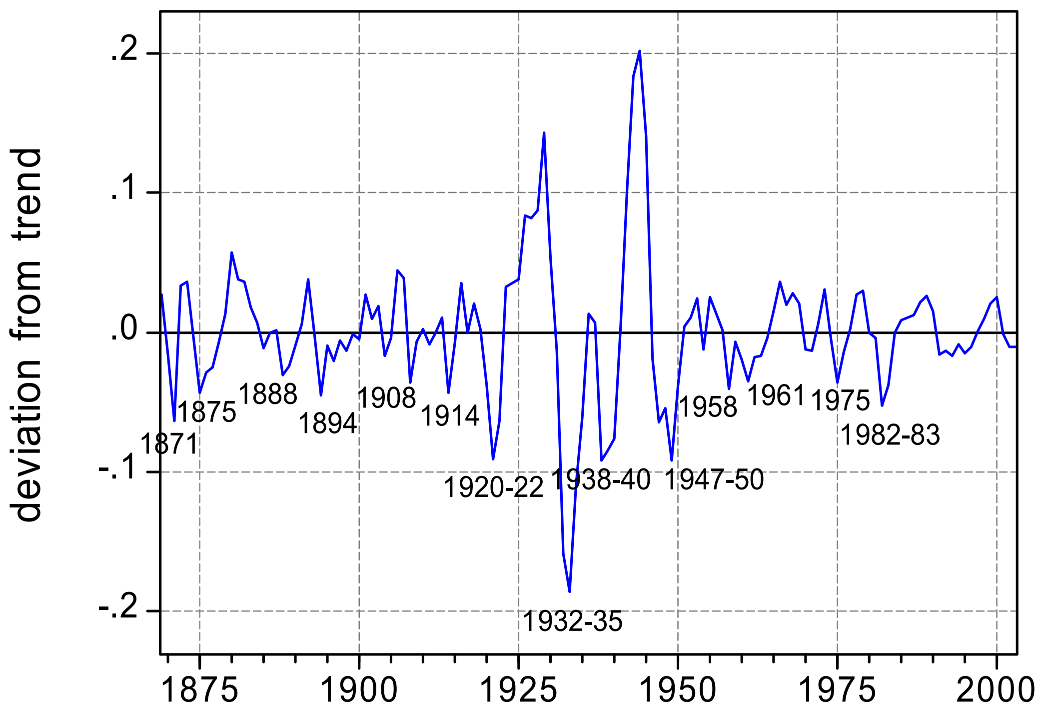


Figure 7.4

Cyclical Part of U.S. Real GDP, 1869-2003

The graph plots the cyclical part of real GDP, which is the difference between the log of real GDP (real GNP before 1929) and its trend. The procedure is analogous to that used in Figures 7.2 and 7.3, except that the underlying data are now annual. The years marked are those for which real GDP fell below its trend by at least 3%.

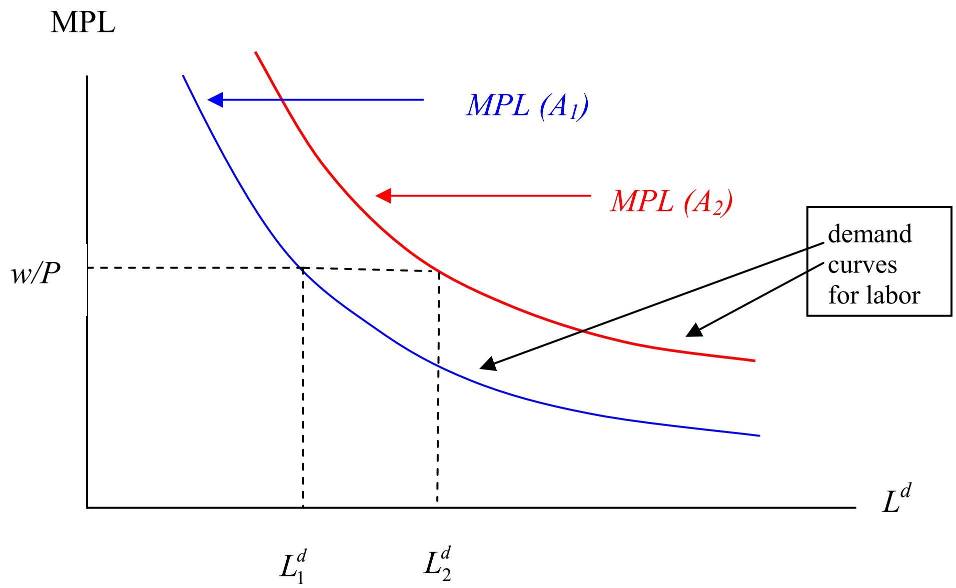


Figure 7.5

Effect of an Increase in the Technology Level on the Demand for Labor

When the technology level is A_1, the marginal product of labor, MPL, is given by the blue curve, labeled as MPL (A_1). This curve is the same as the one in Figure 6.4. At the real wage rate w/P, shown on the vertical axis, the quantity of labor demanded is L_1^d on the horizontal axis. The technology level A_2 is greater than A_1. Therefore, the MPL, given by the red curve labeled MPL (A_2), is higher at any labor input than the value along the blue curve. When the technology level is A_2 and the real wage rate is w/P, the quantity of labor demanded is L_2^d, which is greater than L_1^d.

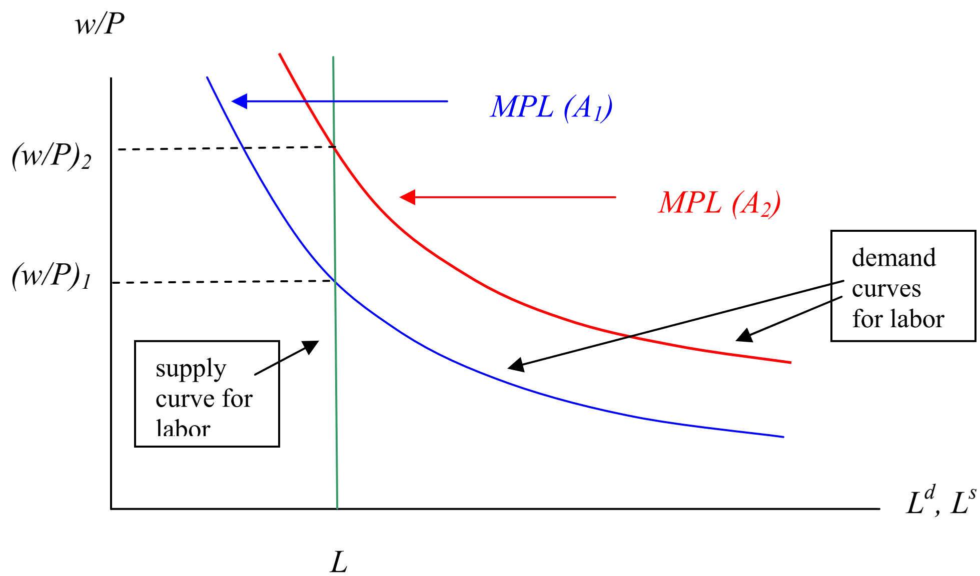


Figure 7.6

Effect of an Increase in the Technology Level on the Real Wage Rate

Labor supply is the given value *L*, shown on the horizontal axis. If the technology level is A_1, the schedule for the MPL determines the blue labor-demand curve, labeled as MPL (A_1). Therefore, the market-clearing real wage rate is $(w/P)_1$, shown on the vertical axis. The technology level A_2 is greater than A_1, as in Figure 7.5. Therefore, the schedule for the MPL is given by the red labor-demand curve, labeled as MPL (A_2). In this case, the market-clearing real wage rate is $(w/P)_2$, which is greater than $(w/P)_1$.

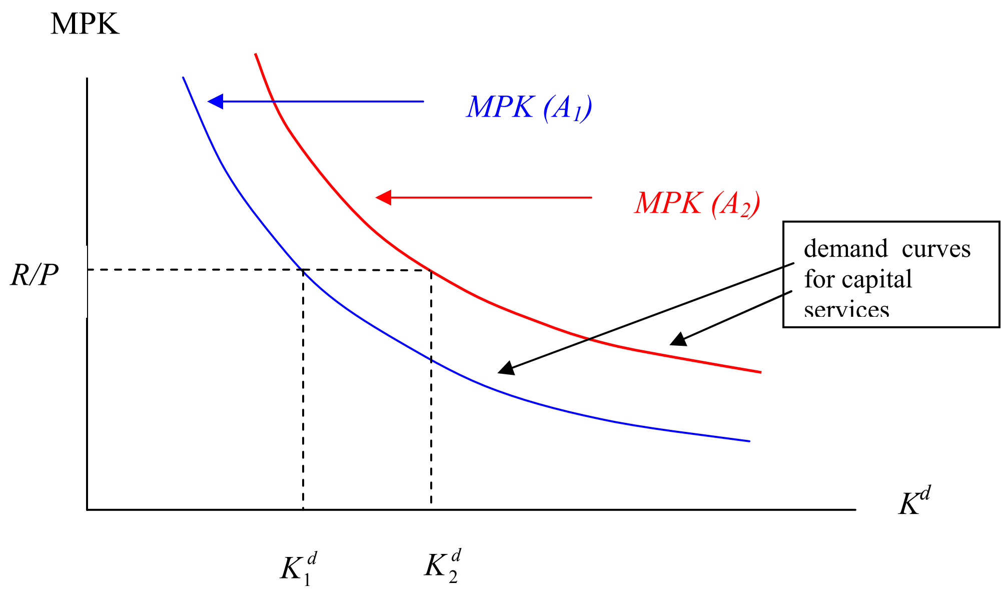


Figure 7.7

Effect of an Increase in the Technology Level on the Demand for Capital Services

When the technology level is A_1, the marginal product of capital, MPK, is given by the blue curve, labeled as MPK (A_1). This curve is the same as the one in Figure 6.6. At the real rental price R/P, shown on the vertical axis, the quantity of capital demanded is K_1^d on the horizontal axis. The technology level A_2 is greater than A_1. Therefore, the MPK, given by the red curve labeled MPK (A_2), is higher at any capital input than the value along the blue curve. When the technology level is A_2 and the real rental price is R/P, the quantity of capital demanded is K_2^d, which is greater than K_1^d.

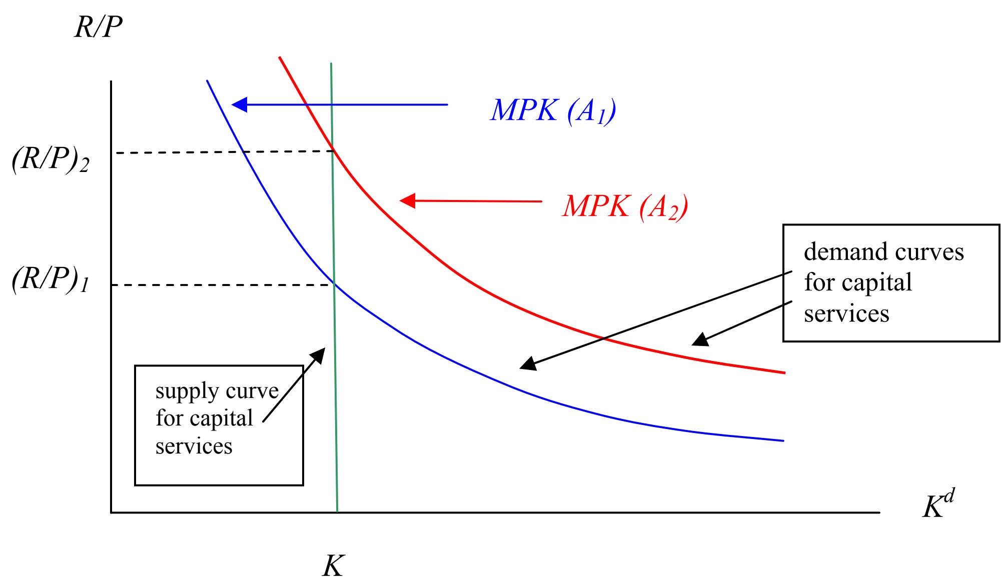


Figure 7.8

Effect of an Increase in the Technology Level on the Real Rental Price of Capital

The supply of capital services is the given value K, shown on the horizontal axis. If the technology level is A_1, the schedule for the MPK is given by the blue curve, labeled as MPK (A_1). This curve gives the demand for capital services when the technology level is A_1. The market-clearing real rental price is $(R/P)_1$, shown on the vertical axis. The technology level A_2 is greater than A_1, as in Figure 7.7. Therefore, the schedule for the MPK is given by the red curve, labeled as MPK (A_2). This curve gives the demand for capital services when the technology level is A_2. In this case, the market-clearing real rental price is $(R/P)_2$, which is greater than $(R/P)_1$.

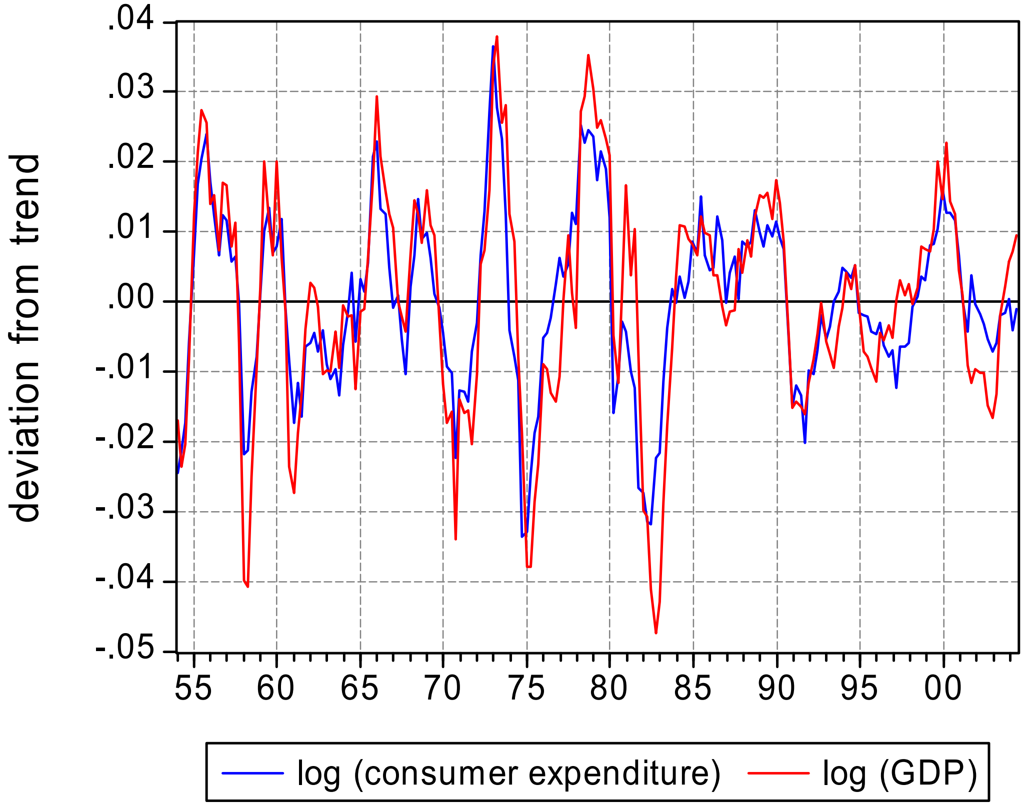


Figure 7.9

Cyclical Parts of U.S. GDP and Consumer Expenditure

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of real consumer expenditure from its trend. The data on GDP and consumer expenditure are quarterly and seasonally adjusted. Real consumer expenditure is procyclical—it fluctuates closely with real GDP but is less variable than real GDP.

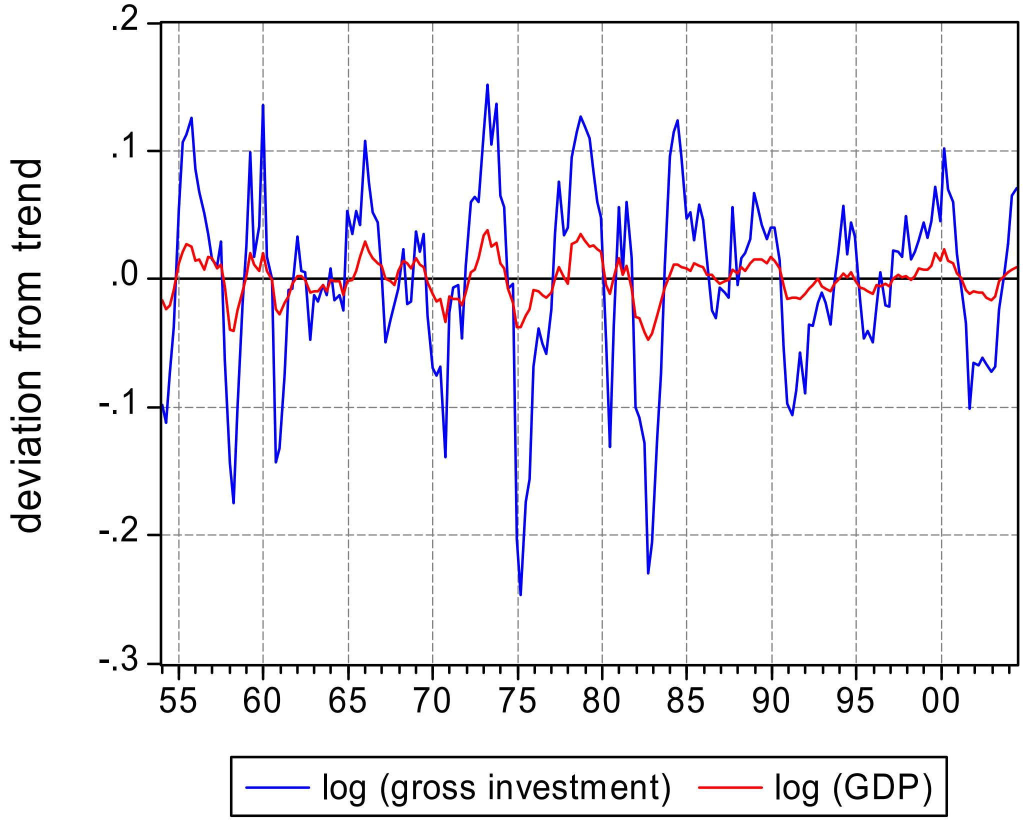


Figure 7.10

Cyclical Parts of U.S. GDP and Investment

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of real gross private domestic investment from its trend. The data on GDP and investment are quarterly and seasonally adjusted. Real gross investment is procyclical—it fluctuates closely with real GDP but is far more variable than real GDP.

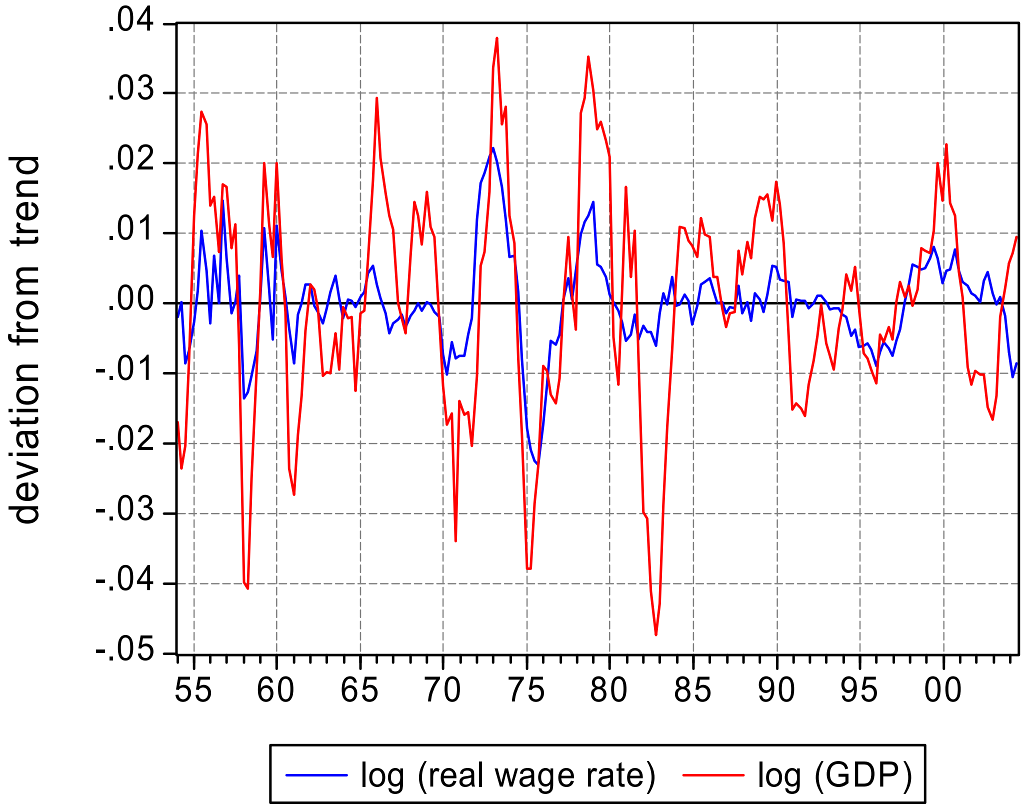


Figure 7.11

Cyclical Parts of U.S. GDP and the Real Wage Rate

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of the real wage rate from its trend. The real wage rate is average hourly nominal earnings of production workers in the total private, non-agricultural economy divided by the price deflator for the GDP. (Before 1964, hourly earnings are for production workers only in manufacturing.) The data on GDP and wage rates are quarterly and seasonally adjusted. (The underlying data on wage rates are monthly.) The real wage rate is procyclical—it fluctuates with real GDP but is not as variable as real GDP.

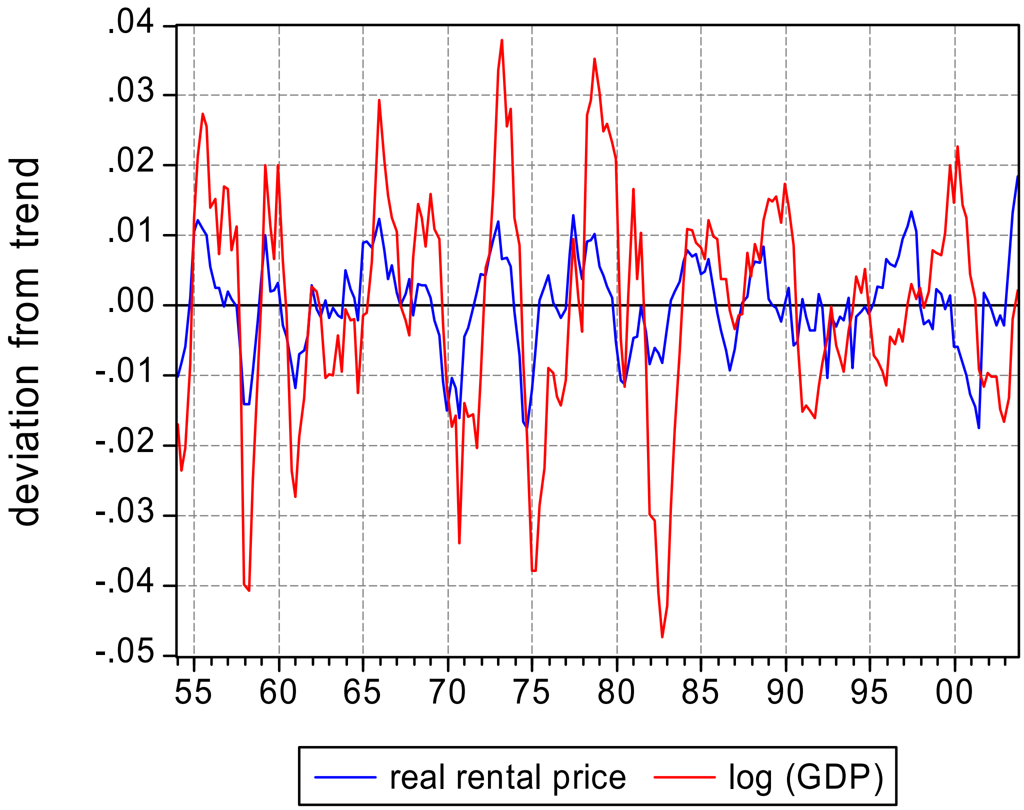


Figure 7.12

Cyclical Parts of U.S. GDP and the Real Rental Price of Capital

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the real rental price of corporate capital from its trend. The real rental price was calculated by Casey Mulligan, based on after-tax payments to capital per unit of capital in the U.S. corporate sector. The real rental price is procyclical—it fluctuates with real GDP.

Appendix

The multi-year budget constraint and the planning horizon

We show here how to calculate the household budget constraint over many years. When we considered two years, we got the budget constraint

(7.7) $C_1 + C_2/(1+i_1) = (1+i_0)\cdot(B_0/P+K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1) - (B_2/P+K_2)/(1+i_1)$.

To extend to many years, we begin with year 3.

The real assets held at the end of year 2 are given by

(7.5) $(B_2/P + K_2) = (1+i_1)\cdot(B_1/P + K_1) + (w/P)_2\cdot L - C_2$.

The real assets held at the end of year 3 are given by an analogous formula, with everything updated by one year:

(7.12) $(B_3/P + K_3) = (1+i_2)\cdot(B_2/P + K_2) + (w/P)_3\cdot L - C_3$.

We found before that we could express the real assets held at the end of year 2 by

(7.6) $(B_2/P + K_2) = (1+i_1)\cdot(1+i_0)\cdot(B_0/P + K_0) + (1+i_1)\cdot (w/P)_1\cdot L - (1+i_1)\cdot C_1 + (w/P)_2\cdot L - C_2$.

If we replace $(B_2/P + K_2)$ on the right-hand side of equation (7.7) by the right-hand side of equation (7.6), we get

$$(B_3/P+K_3) = (1+i_2)\cdot\left[(1+i_1)\cdot(1+i_0)\cdot(B_0/P+K_0)+(1+i_1)\cdot(w/P)_1\cdot L-(1+i_1)\cdot C_1+(w/P)_2\cdot L-C_2\right]$$
$$+(w/P)_3\cdot L-C_{3.}$$

If we multiply through $1+i_2$ by the terms inside the brackets, we get

$$(7.13)\quad (B_3/P + K_3) = (1+i_2)\cdot(1+i_1)\cdot(1+i_0)\cdot(B_0/P + K_0) + (1+i_2)\cdot(1+i_1)\cdot(w/P)_1\cdot L$$
$$- (1+i_2)\cdot(1+i_1)\cdot C_1 + (1+i_2)\cdot(w/P)_2\cdot L - (1+i_2)\cdot C_2 + (w/P)_3\cdot L - C_3.$$

The important result involves the interest-rate terms. The initial real assets, $(B_0/P + K_0)$, now accumulate interest over three years, up to the end of year 3. Thus, these assets are multiplied by $(1+i_2)\cdot(1+i_1)\cdot(1+i_0)$. Year 1's real wage income, $(w/P)_1\cdot L$, accumulates interest over two years and is therefore multiplied by $(1+i_2)\cdot(1+i_1)$. The other real income and consumption terms enter in an analogous way.

If we divide through everything in equation (7.13) by $(1+i_2)\cdot(1+i_1)$ and rearrange the terms to put only those involving consumption on the left-hand side, we get the **3-year budget constraint**:

$$(7.14) \quad C_1 + C_2/(1+i_1) + C_3/[(1+i_1)\cdot(1+i_2)] = (1+i_0)\cdot(B_0/P+K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1) + (w/P)_3\cdot L/[(1+i_1)\cdot(1+i_2)] - (B_3/P+K_3)/[(1+i_1)\cdot(1+i_2)].$$

This result extends the two-year budget constraint from equation (7.7) to three years. Everything in equation (7.14) appears as a present value (or year 1 value). But now the budget constraint includes the real wage income and consumption from year 3, $(w/P)_3\cdot L$ and C_3, and these amounts are discounted for the accumulation of interest over two years, that is, by $(1+i_1)\cdot(1+i_2)$. The real assets held at the end of year 3, $(B_3/P + K_3)$, now appear on the right-hand side and are also discounted by $(1+i_1)\cdot(1+i_2)$.

By now, we see how to extend the budget constraint to any number of years. Each time we push forward one more year, we bring in the real income and consumption from that year. We also bring in the real assets held at the end of the new year and drop the real assets held at the end of the previous year. All of the new terms are discounted to reflect the accumulation of interest from year 1 up to the future year. For example, if we consider j years, where j is greater than 3, we get the **j-year budget constraint**:

$$(7.15) \quad C_1 + C_2/(1+i_1) + C_3/[(1+i_1)\cdot(1+i_2)] + \ldots + Cj/[(1+i_1)\cdot(1+i_2)\cdot \ldots \cdot(1+i_{j-1})] = (1+i_0)\cdot(B_0/P + K_0) + (w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1) + (w/P)_3\cdot L/[(1+i_1)\cdot(1+i_2)] + \ldots + (w/P)_j\cdot L/[(1+i_1)\cdot(1+i_2)\cdot \ldots \cdot(1+i_{j-1})] - (B_j/P + K_j)/[(1+i_1)\cdot(1+i_2)\cdot \ldots \cdot(1+i_{j-1})].$$

We want to use equation (7.15) to understand how households choose year 1's consumption, C_1. The idea is that households make this choice as part of a long-term plan that considers future consumptions and real incomes. These future values relate to the current choice through the j-year budget constraint. We can think of the number j as the **planning horizon**.

How long is the horizon that households consider when making decisions? Because we are dealing with households that have access to borrowing and lending, a long horizon is appropriate. By borrowing or lending, households can effectively use future income to pay for current consumption, or current income to pay for future consumption. When expressed as a present value, future incomes are as pertinent for current decisions as is today's income.

Economists often assume that the planning horizon is long but finite. For example, in theories called **life-cycle models**, the horizon, j, represents an individual's expected remaining lifetime.[17] If people do not care about things that occur after their death, they have no reason to carry assets beyond year j. Accordingly, they would plan so as to set to zero the final stock of assets, $(B_j/P + K_j)$, in equation (7.15).[18] That is, each person plans to end up with no assets when he or she dies.

It is straightforward to define the anticipated lifetime and, thereby, the planning horizon for an isolated individual. The appropriate horizon is, however, not so obvious for a family in which individuals have spouses and children. Since people care about their spouse and children, the applicable horizon extends beyond a person's expected

[17] Life-cycle models are associated particularly with the economist Franco Modigliani. See Modigliani and Brumberg (1954) and Ando and Modigliani (1963).
[18] We have to rule out the possibility of dying with negative assets. Otherwise, households would like to set $(B_j/P + K_j)$ to be a large negative number.

lifetime, and households would give weight to the expected future real income and consumption of their children. Further, since children care about their prospective children, should they have any, there is no clear point at which to draw the line.

Instead of imposing a **finite horizon**, we can think of households' plans as having an **infinite horizon**. There are two good reasons to make this assumption:

- First, if we think of the typical person as part of a family that cares about the members of future generations—children, grandchildren, and so on—into the indefinite future, this setup is the correct one.
- Second, although it is not obvious at this point, an infinite horizon is the simplest assumption.

If we use an infinite horizon, we allow the number j to become arbitrarily large in equation (7.15). In that case, there is no final year, j, and we do not have to worry about the last term on the right-hand side, which involves $(B_j/P + K_j)$.[19] Thus, the **infinite-horizon budget constraint** is the one we used before:

$$(7.10) \qquad C_1 + C_2/(1+i_1) + C_3/[(1+i_1)\cdot(1+i_2)] + \ldots = (1+i_0)\cdot(B_0/P + K_0)$$
$$+ (w/P)_1\cdot L + (w/P)_2\cdot L/(1+i_1) + (w/P)_3\cdot L/[(1+i_1)\cdot(1+i_2)] + \ldots$$

The ellipses signify that we are including terms involving C_t and $(w/P)_t\cdot L$ out into the indefinite future.

[19]Because of the discounting by $(1+i_1)\cdot(1+i_2)\cdot \ldots \cdot(1+i_{j-1})$, the present value of the assets left over, $(B_j/P + K_j)$, tends to become negligible as j gets very large.

Chapter 8

Labor Supply and Capital Utilization

Thus far, we assumed that household labor supply was a constant, $L^s = L$. The real wage rate, w/P, adjusted to equate labor demand, L^d, to L^s—therefore, the quantity of labor input always equaled the fixed amount L. We also assumed that households supplied all of their capital to the rental market, so that $K^s = K$, where K is the given stock of capital. The real rental price, R/P, adjusted to equate the demand for capital, K^d, to K^s—therefore, the capital stock, K, was always fully utilized in production. Or, to put in another way, the **capital utilization rate** was always 100%.

Since real GDP is given by

$$Y = A \cdot F(K,L), \tag{3.1}$$

the values of K and L determined real GDP, for a given technology level, A. Therefore, with K and L fixed, real GDP changed only because of changes in A.

Two shortcomings of this analysis are the failures to fit the observed patterns for labor input and the capital utilization rate during economic fluctuations. Later in this chapter, we shall look in detail at how these variables behave during booms and recessions. However, to motivate our theoretical analysis in this chapter, we can note that labor input—measured by employment or total hours worked—and the capital utilization rate are strongly procyclical. These macroeconomic variables are high in booms and low in recessions. We cannot give our model high grades unless we can extend it to explain these important phenomena.

To fit the facts on labor input and capital utilization, we now extend the model to allow for variability of labor supply, L^s, and the supply of capital services, K^s. These extensions will be important for two reasons. First, we will be able to explain short-term variations in labor input and the capital utilization rate. Second, changes in real GDP will reflect short-term variations in labor and capital input, as well as changes in the technology level, A.

I. Labor Input

We begin by extending the microeconomic foundations to allow for variable labor supply. Then we use our market-clearing approach to the labor market to assess the determination of the quantity of labor input.

A. Labor supply

We begin with a modified form of the household budget constraint that we worked out in equation (7.1) of chapter 7:

$$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L^s + i\cdot(B/P + K). \tag{8.1}$$

The left-hand side is the total of consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$. The right-hand side is household real income, which is the sum of real wage income, $(w/P)\cdot L^s$, and real income on assets, $i\cdot(B/P + K)$. The difference from before is that we replaced L by the quantity of labor supplied, L^s, to allow for a variable labor supply.

Since each household has a fixed amount of time available each year, a higher quantity of labor supplied, L^s, means a smaller amount of leisure.[1] We have already assumed that households like higher consumption, C, at each point in time. Now we assume that households also like more leisure at each point in time. That is, leisure can be thought of as another form of consumption. To put it a different way, households dislike higher work effort, represented by labor supply, L^s.

As with consumer demand and saving, the choice of L^s will involve substitution and income effects. We start with the substitution effect for leisure and consumption.

1. The substitution effect for leisure and consumption. Consider the household budget constraint in equation (8.1). The right-hand side includes the real wage rate, w/P, and the interest rate, i, each of which an individual household takes as given. Suppose that we also hold fixed the real assets, $B/P + K$, on the right-hand side, and the real saving, $(1/P)\cdot\Delta B + \Delta K$, on the left-hand side. In this case, households can choose to raise or lower the quantity of labor supplied, L^s, and thereby raise or lower real wage income, $(w/P)\cdot L^s$. Since we are holding everything else fixed in equation (8.1), the higher or lower real wage income must be used to raise or lower consumption, C. In other words, if households choose to work one more hour and thereby have one hour less leisure, the extra w/P of real wage income pays for w/P more units of consumption. Therefore, the household can substitute one less hour of leisure for w/P more units of consumption.

[1] From the perspective of a household, a higher L^s can mean either that family members who are working work more hours per year or that more of the members work. In the latter case, the rise in labor supply shows up as an increase in labor-force participation. Either way, more labor supplied means less leisure for the family as a whole.

If the real wage rate, w/P, rises, households get a better deal by working more. That is, they get more consumption for each extra hour worked and, hence, for each hour of leisure given up. Since the deal is better, we predict that households respond to a higher w/P by taking more advantage of the deal—they work more, have more real wage income and consumption, and enjoy less leisure. Another way to say the same thing is that a higher real wage rate, w/P, makes leisure more expensive compared to consumption. The reason is that w/P tells households how much consumption they give up by taking an extra hour of leisure (and working less). An increase in w/P motivates households to substitute away from the thing that got more expensive—leisure—and toward the thing that got cheaper—consumption. Therefore, a higher real wage rate, w/P, raises the quantity of labor supplied, L^s.

2. Income effects on labor supply. As usual, we also have to consider income effects. We can see from the budget constraint,

$$C + (1/P)\cdot \Delta B + \Delta K = (w/P)\cdot L^s + i\cdot (B/P + K), \tag{8.1}$$

that a change in the real wage rate, w/P, has an income effect. Specifically, for a given quantity of labor supplied, L^s, a higher w/P means higher real wage income, $(w/P)\cdot L^s$. Our prediction is that households spend the extra income on the two things that they like—consumption and leisure. Thus, on this ground, a higher w/P leads to more leisure and, hence, lower labor supply, L^s. Since we already know that the substitution effect from a higher w/P favors higher L^s, the overall effect is ambiguous. An increase in the real wage rate, w/P, raises L^s if the substitution effect is stronger than the income effect but lowers L^s if the substitution effect is weaker than the income effect.

We can say more by considering cases where the income effect is strong or weak. We found in chapter 7 that the strength of the income effect depended on whether a change was permanent or temporary. To see how this works, we can use a modified form of the multi-year budget constraint that we worked out in equation (7.10):

$$(8.2) \quad C_1 + C_2/(1+i_1) + C_3/[(1+i_1)\cdot(1+i_2)] + \ldots = (1+i_0)\cdot(B_0/P + K_0)$$

$$+ (w/P)_1\cdot L_1^s + (w/P)_2\cdot L_2^s/(1+i_1) + (w/P)_3\cdot L_3^s/[(1+i_1)\cdot(1+i_2)] + \ldots$$

The difference from before is that we replaced the fixed quantity of labor, L, by the quantity of labor supplied in each year, L_t^s.

We considered before the income effects on consumer demand. We argued that households responded to higher real wage rates—and, hence, higher real wage incomes—by consuming more each year. That is, the income effect was positive for each year's consumption. However, the income effect was much stronger if the change in the real wage rate was permanent—applying to $(w/P)_2$, $(w/P)_3$, ...—rather than just to the current year, $(w/P)_1$.

The same reasoning holds for the income effect on labor supply. If an increase in this year's real wage rate, $(w/P)_1$, is permanent—so that $(w/P)_2$, $(w/P)_3$, ... also rise—the income effect is large. In this case, we are unsure whether the full effect from an increase in year 1's real wage rate, $(w/P)_1$, is to raise or lower year 1's labor supply, L_1^s. That is, the income effect, which lowers labor supply, may be stronger or weaker than the substitution effect, which raises labor supply.

In contrast, if the change in year 1's real wage rate, $(w/P)_1$, is temporary—so that $(w/P)_2$, $(w/P)_3$, ... do not change—the income effect is small. In this case, we are pretty sure that the income effect will be weaker than the substitution effect. Therefore, a temporary increase in year 1's real wage rate, $(w/P)_1$, would raise year 1's labor supply, L_1^s.

3. Intertemporal-substitution effects on labor supply. We found before that changes in interest rates, i, had intertemporal-substitution effects on consumption and saving. Now we can use equation (8.2) to study intertemporal-substitution effects on labor supply.

We found before that an increase in year 1's interest rate, i_1, made year 2's consumption, C_2, cheaper compared with year 1's, C_1. This effect can be seen in equation (8.2), which shows that C_2 is discounted by $1+i_1$ to get a present value before combining it with C_1. From this perspective, we predicted that an increase in i_1 tended to lower C_1 and raise C_2.

Year 2's real wage income, $(w/P)_2 \cdot L_2^s$, is also discounted by $1+i_1$ to get a present value before combining it with year 1's, $(w/P)_1 \cdot L_1^s$. Therefore, if the interest rate, i_1, rises, we predict that households would increase L_1^s and decrease L_2^s. This intertemporal-substitution effect from a higher interest rate favors more labor supply today and less in the future.

Another way to say the same thing is that a higher interest rate, i_1, motivates households to reduce current leisure and raise future leisure. The reasoning is the same as for consumption—a higher i_1 means that future consumption and leisure have become cheaper relative to current consumption and leisure. Therefore, households substitute toward the cheaper things—future consumption and leisure—and away from the more expensive things—current consumption and leisure.

Now we consider intertemporal-substitution effects from variations of the real wage rate over time. Suppose that we start from equal real wage rates in each year. Assume then that year 1's real wage rate, $(w/P)_1$, rises, while future real wage rates—$(w/P)_2$, $(w/P)_3$, ...—do not change. In this case, households are motivated to supply a lot of labor when the real wage rate is temporarily high—year 1—and relatively little labor in subsequent years.[2] Hence, a rise in year 1's real wage rate, $(w/P)_1$, raises year 1's labor supply, L_1^s, because of this intertemporal-substitution effect.

To put the results together, assume first that the rise in year 1's real wage rate, $(w/P)_1$, is temporary. In this case, the forces that favor higher L_1^s are the substitution effect between consumption and leisure and the strong intertemporal-substitution effect. The force that favors lower L_1^s is the weak income effect. Since the income effect is weak, we can be confident that L_1^s rises overall.

The conclusions are different if the rise in year 1's real wage rate, $(w/P)_1$, is permanent. In this case, the forces that favor higher L_1^s are the substitution effect between consumption and leisure and the weak intertemporal-substitution effect. The force that favors lower L_1^s is the strong income effect. Since the income effect is strong, we are now uncertain whether L_1^s rises or falls overall.

In our analysis in this chapter, it will be important to get the "right slope" for the labor-supply function—that is, the analysis will rely on the assumption that the overall effect from an increase in the real wage rate, $(w/P)_1$, is to raise the current quantity of labor supplied, L_1^s. We can see from the discussion that this property is more likely to

[2] Another way to say the same thing is that current leisure is relatively cheap when the real wage rate is temporarily low. Therefore, a period of temporarily low real wage rates is a good time to take a vacation.

hold if the rise in w/P is less than fully permanent. We shall assume in this chapter, as in chapter 7, that the underlying sources of economic disturbances are changes in the technology level, A—or other disturbances, such as oil shocks and harvest failures, that resemble changes in A. The resulting changes in w/P will be less than fully permanent if the changes in A are themselves less than fully permanent. Thus, to get the right slope of the labor-supply function, we will want to assume that the shocks to A are less than fully permanent. This conclusion reinforces our reasoning from chapter 7. We argued there that, to explain why investment was far more variable than consumption, it helped to assume that the shocks to A were typically less than fully permanent.

Empirical evidence on intertemporal substitution of labor supply

Macroeconomists usually emphasize the intertemporal-substitution effect of interest rates on consumption and saving. However, there is evidence that interest rates and time-varying real wage rates also have intertemporal-substitution effects on labor supply.

George Alogoskoufis (1987b) found for U.S. household data from 1948 to 1982 that an increase in the expected growth rate of real wage rates by one percentage point per year raised the growth rate of employment (number of persons employed) by about one percentage point per year. Thus, employment was deferred when workers thought that future real wage rates would be higher than current real wage rates. An increase in the annual interest rate by one percentage point lowered the growth rate of employment by 0.6 percentage point per year. Thus, a higher interest rate motivated more work today and less work in the future. For British household data from 1950 to 1982, the effects (reported in Alogoskoufis [1987a]) were smaller but still statistically significant. An increase in the expected growth rate of real wage rates by one percentage point per year raised the growth rate of employment by 0.4 percentage point. An increase in the interest rate by one percentage point lowered the growth rate of employment by 0.2 percentage point per year. Alogoskoufis did not find any significant effects of time-varying real wage rates or interest rates on the growth rate of hours worked per person. Therefore, his findings suggest that intertemporal substitution of labor supply is more important for employment than for hours worked per person. This pattern is surprising—one would have thought that workers would substitute intertemporally by working extra hours, including overtime and weekends, when real wage rates were thought to be temporarily high.

Casey Mulligan (1995b) argued that it is hard to detect intertemporal-substitution effects from time-varying real wage rates in the household data used by Alogoskoufis. The problem is that it is unclear when households perceive their current real wage rates to be temporarily high or low. Therefore, Mulligan looked at unusual events for which the temporary nature of high real wage rates was clear. One example is the construction of the Alaskan gas pipeline from 1974 to 1977. Mulligan found that the temporarily high real wage rates paid to construction workers elicited a substantial increase in labor supply. A temporary rise of real wage rates by 10% was estimated to increase average hours worked per week by about 20%. A second example is the cleanup of the *Exxon Valdez* oil spill in Alaska in 1989. Mulligan found that the temporarily high real wage rates paid to workers in the transportation and public utilities industries induced a sharp increase in labor supply. His estimate was that a temporary rise in real wage rates by 10% raised average hours worked per week by even more than 20%. Thus, unlike Alogoskoufis, Mulligan found a substantial response of hours worked per person to temporary changes in real wage rates. The evidence from the two Alaskan "experiments" is that workers are willing to expand labor supply dramatically in response to temporary opportunities for high real wage rates.

B. Labor input during economic fluctuations

In chapter 7, we assumed that market labor supply, L^s, was fixed each year at L. Therefore, fluctuations of real GDP were not accompanied by changes in labor input. Since we now allow labor supply to be flexible, we can get changes in labor input. Thus, the model has some chance of matching the data, which show that labor input moves in the short run in the same direction as real GDP, that is, labor input is procyclical.

1. The cyclical behavior of labor input: empirical patterns. To see what we are trying to explain, Figures 8.1 and 8.2 use information on two concepts of labor input for the United States. The first is **employment** (numbers of persons with jobs) and the second is **total hours worked** (which multiplies employment by the average weekly hours worked per employee).[3] In each case, we use the method from Figure 7.2 to detrend the log of the variable. Then we calculate the deviations of the log of each variable from its trend—as before, these are our measures of cyclical parts. We show the cyclical parts of employment and total hours worked as the blue graphs in Figures 8.1 and 8.2, respectively. As before, the red graphs show the cyclical parts of real GDP.

We see from the graphs that both measures of labor input are strongly procyclical—they move in the same direction as real GDP.[4] The variability of labor input is nearly as great as that of real GDP—the standard deviations of the cyclical parts are 1.3% for employment and 1.5% for total hours, compared with 1.6% for real GDP.

2. The cyclical behavior of labor input: theory. Now we want to see whether the model accords with the observed behavior of labor input. In our earlier analysis, we matched some features of economic fluctuations by considering shocks to the technology

[3] The employment numbers come from the establishment survey of the Bureau of Labor Statistics (BLS). These numbers exclude agricultural employment, self-employment, and a few other types of labor input. Total hours worked is the product of employment and average weekly hours in the total private economy (multiplied by 52 weeks per year). These data on average hours worked are available from the BLS only since 1964.

[4] From 1954.1 to 2004.3, the correlation of the cyclical part of employment with the cyclical part of real GDP is 0.79. From 1964.1 to 2004.3, the correlation of the cyclical part of total hours worked with the cyclical part of real GDP is 0.88.

level, A. The model worked best when the changes in A were long-lasting but not permanent. For fixed factor inputs, K and L, a rise in A led directly to an increase in real GDP, $A \cdot F(K,L)$. Some of this increase showed up as more consumption and some as more investment. The change also raised the marginal products of labor, MPL, and capital, MPK. Therefore, the real wage rate, w/P, the real rental price of capital, R/P, and the interest rate, i, all increased. Now we want to see how an increase in A affects labor input, L, when we allow labor supply, L^s, to vary.

a. Variations in labor supply. The substitution effect between consumption and leisure implies that the higher real wage rate, w/P, motivates households to increase consumption, reduce leisure, and raise labor supply. The important point is that this substitution effect raises the quantity of labor supplied, L^s.

Next are the income effects. We see from equation (8.2) that the increases in the real wage rate, w/P, and the interest rate, i, raise household income. Moreover, since we are considering a long-lasting increase in A, the income effects are strong—hence, consumption and leisure both tend to rise. The increase in leisure means that L^s declines—this response works against the substitution effect between consumption and leisure.

Finally, consider the intertemporal-substitution effects. The increase in the interest rate, i, motivates substitution of labor supply toward the present and away from the future. Thus, this effect tends to raise current labor supply, L^s—that is, it reinforces the substitution effect between consumption and leisure.[5] Since the change in A is less

[5] However, the increase in i also has an income effect, which tends to raise leisure and reduce labor supply, L^s.

than permanent, the increase in the real wage rate, w/P, will also not be permanent. Since w/P is temporarily high, we have another intertemporal-substitution effect that tends to raise current labor supply, L^s.

Overall, the change in current labor supply, L^s, reflects the offset of substitution effects—which favor higher labor supply—against income effects—which favor lower labor supply. To match the procyclical behavior of labor input shown in Figures 8.1 and 8.2, we must have that the substitution effects dominate, so that market labor supply, L^s, rises overall. As mentioned before, this "right slope" for the labor-supply function is more likely to hold the less persistent are the underlying shocks to A. When the shifts to A are temporary, the changes in w/P are also temporary, and the income effects are relatively weak.

b. Labor demand and the labor market. Figure 8.3 shows how an increase in the technology level, A, affects the labor market. The downward-sloping labor-demand curves are the same as those shown in Figure 7.5. The blue curve shows market labor demand, L^d, at an initial technology level, A_1. The curve slopes downward, because a decrease in the real wage rate, w/P, raises the quantity of labor demanded. The downward-sloping red curve applies for a higher technology level, A_2. Since A_2 is greater than A_1, the marginal product of labor, MPL, is higher at any given L along the red curve than along the blue one. Therefore, the quantity of labor demanded, L^d, is higher at any w/P along the red curve. Thus, the red curve, labeled L^d (A_2), lies to the right of the blue curve, labeled L^d (A_1).

Figure 8.3 also shows an upward-sloping curve for labor supply, L^s. This curve slopes up because we assume that the substitution effect from a higher real wage rate, *w/P*, dominates over the income effect. We assume that the same curve for L^s applies for the two technology levels. That is, for a given *w/P*, we assume that the labor-supply curve does not shift when the technology level rises from A_1 to A_2.[6]

We get two important conclusions from Figure 8.3. First, as in chapter 7, the real wage rate rises, from $(w/P)_1$ to $(w/P)_2$. Second, aggregate labor input increases, from L_1 to L_2. The second effect is new, and it depends on the upward slope of the labor-supply curve, L^s, in the graph. Thus, if a higher *w/P* induces greater labor supply, an increase in the technology level raises labor input, *L*. Hence, the model can match the finding in Figures 8.1 and 8.2 that labor input, measured by employment or total hours worked, moves along with real GDP during economic fluctuations.

The increase in labor input also contributes to the rise in real GDP, which equals $A \cdot F(K,L)$. We now have that real GDP rises partly because of the direct effect from a higher technology level, *A*, and partly because of the induced increase in labor input, *L*.

3. The cyclical behavior of labor productivity. Another macroeconomic variable that economists consider is labor productivity. The usual definition of labor productivity—for example, the one used in popular media—is the average product of labor, which is the ratio of real GDP, *Y*, to labor input, *L*. We show in the first appendix to this chapter that the model implies that this concept of labor productivity is procyclical—high in booms and low in recessions. The reason is that the average product

[6] We are neglecting here effects from changes in the interest rate, *i*. A higher *i* raises L^s because of an intertemporal-substitution effect but lowers L^s because of an income effect.

of labor, Y/L, tends to move in the same direction as the marginal product of labor, MPL. We already know that the MPL—which equals the real wage rate, w/P, when the labor market clears—is procyclical.

The appendix looks also at the behavior of labor productivity in the U.S. data. To calculate labor productivity, we measure labor input, L, either by employment or total hours worked. Thus, the first measure of labor productivity, Y/L, is real GDP per worker and the second is real GDP per worker-hour. Labor productivity turns out to be procyclical in both cases. From 1954.1 to 2004.3, the correlation of the cyclical part of real GDP per worker with the cyclical part of real GDP is 0.57. For the second concept of labor productivity—real GDP per worker-hour—we have good data only since 1964.1. From 1964.1 to 2004.3, the correlation of the cyclical part of real GDP per worker-hour with the cyclical part of real GDP is 0.31. Thus, the model fits the broad patterns in the labor-productivity data. However, there may be a puzzle about why real GDP per worker-hour is only weakly procyclical.

II. Capital Input

Now we extend the microeconomic foundations to allow for a variable capital utilization rate and, hence, for a variable supply of capital services. Then we use our market-clearing approach to the rental market to assess the determination of the quantity of capital services.

Up to now, we did not distinguish the stock of capital, K, from the quantity of capital services used in production. That is, if we think of K as the number of machines, we assumed that each machine was utilized a fixed number of hours per year. For example, if businesses use each machine 8 hours per day, 5 days per work, 52 weeks per year, then each machine yields 2080 machine-hours per year. Therefore, capital services—measured as machine-hours per year—would always be a fixed multiple, in this case 2080, of the capital stock.

In practice, the capital utilization rate can vary. For example, if businesses operate each machine 16 hours per day, corresponding to two 8-hour shifts each weekday instead of one, then each machine would yield 4160 machine-hours per year, rather than 2080. Similarly, businesses can raise the utilization rate by operating on weekends.

Let the variable κ (the Greek letter kappa) represent the utilitization rate for the capital stock, K. We measure κ in units of hours per year, and K as the number of machines (or quantity of goods). The product of κ and K, κK, then represents the flow of capital services. The term κK has units of

(hours per year) ▪ (number of machines) = machine-hours per year.

We now modify the production function from equation (3.1) to replace the capital stock, K, by the quantity of capital services, κK:[7]

Key equation (production function with variable capital utilization):

$$Y = A \cdot F(\kappa K, L). \tag{8.3}$$

For given K, κK, rises with the utilization rate, κ. Therefore, an increase in κ raises real GDP, Y, for a given technology level, A, capital stock, K, and labor input, L.

[7] Our assumption is that production depends only on the quantity of capital services per year, κK, and not on how these services break down into the utilization rate, κ, and the number of machines, K. For example, machines operated at night are assumed to be just as productive as machines operated during the day.

A. The demand for capital services

In chapter 6, we worked out the demand for capital services, K^d, as an input to production. Households, as managers of family businesses, chose K^d to maximize real profit:

(6.13) $$\Pi/P = A\cdot F(K^d, L^d) - (w/P)\cdot L^d - (R/P)\cdot K^d.$$

We found that the maximization of Π/P led to the equation of the marginal product of capital, MPK, to the real rental price of capital, R/P. We also found that an increase in R/P reduced the quantity of capital services demanded, K^d, as shown by the downward-sloping curve in Figure 6.6.

This analysis is still valid if we revise equation (6.13) to allow for a variable capital utilization rate, κ:

(8.4) $$\Pi/P = A\cdot F[(\kappa K)^d, L^d)] - (w/P)\cdot L^d - (R/P)\cdot(\kappa K)^d.$$

In this new formulation, the real rental price, R/P, is measured per unit of capital services. That is, since κK has the units of machine-hours per year, R/P has the units of goods per machine-hour.

Households again choose the demand for capital services, now represented by $(\kappa K)^d$, to maximize real profit, Π/P. The maximization still entails the equation of the marginal product, MPK, to the real rental price, R/P. However, we should now think of MPK as the additional goods produced by an extra machine-hour of capital services. The resulting demand curve for capital services, $(\kappa K)^d$, still looks like the one shown in Figure 6.6. We show this demand as the downward-sloping blue curve in Figure 8.4.

Assume now that the technology level rises from A_1 to A_2. This change raises the marginal product of capital services, MPK, at a given quantity of capital services, κK.

Figure 8.5 shows this shift as a movement from the blue curve to the red one. Thus, at the given real rental price, *R/P*, the demand for capital services rises from $(\kappa K)_1^d$ to $(\kappa K)_2^d$. These results are similar to those found before in Figure 7.6, where we did not allow for a variable capital utilization rate.

B. The supply of capital services

On the supply side, we assumed in chapter 6 that owners of capital (households) supplied all of their capital, *K*, to the rental market. However, owners now have a choice about the capital utilization rate, *κ*. For a given stock of capital, *K*, owners can supply more or less capital services per year by varying *κ*. The question is why would an owner ever set *κ* below its maximum possible value? This maximum rate, corresponding to the operation of machines 24 hours per day and 7 days per week, is 8736 hours per year.

One reason to set the utilization rate, *κ*, at less than its maximum value is that increases in *κ* tend to raise the depreciation rate, *δ*. Machines wear out faster if they are used more intensively. Moreover, as *κ* rises, the time available for maintenance declines, thereby contributing further to the higher depreciation rate. We can capture these effects by writing the depreciation rate as an upward-sloping function of *κ*:[8]

$$\delta = \delta(\kappa).$$

Owners of capital choose the utilization rate, *κ*, to maximize the net real income from supplying capital services:

$$\begin{aligned} \textit{net real income from supplying capital services} &= \textit{real rental payments - depreciation} \\ &= (R/P)\cdot\kappa K - \delta(\kappa)\cdot K. \end{aligned}$$

[8] For a formal model that introduced this idea, see Jeremy Greenwood, Zvi Hercowitz, and Gregory Huffman (1988).

If we take outside the variable K, we can write the result as

$$(8.5) \qquad net\ real\ income\ from\ supplying\ capital\ services = K\cdot[(R/P)\cdot\kappa - \delta(\kappa)].$$

Thus, the net real income equals the capital owned, K, multiplied by the term $(R/P)\cdot\kappa - \delta(\kappa)$. The first term in the last expression, $(R/P)\cdot\kappa$, is the real rental income per year on each unit of capital. Note that this expression is the product of the real rental per machine-hour, R/P, and the machine-hours per year, κ, provided by each machine. When we subtract the depreciation rate, $\delta(\kappa)$, we end up with the net real rental income per unit of capital, $(R/P)\cdot\kappa - \delta(\kappa)$. This expression also equals the rate of return from owning capital:

Key equation (rate of return on capital):

$$(8.6) \qquad rate\ of\ return\ from\ owning\ capital = (R/P)\cdot\kappa - \delta(\kappa).$$

For a given capital stock, K, the maximization of the net real income from supplying capital services boils down to maximizing the net real return from owning capital, $(R/P)\cdot\kappa - \delta(\kappa)$. Figure 8.6 graphs the two parts of this return against the utilization rate, κ. The green straight line shows the first part, $(R/P)\cdot\kappa$. This line starts from the origin and has a slope equal to the real rental price, R/P. As in our previous analysis, an individual household takes R/P as given.

The second part of the net real income per unit of capital is the negative of the depreciation rate, $\delta(\kappa)$. We graph $\delta(\kappa)$ versus κ in Figure 8.6 as the blue curve. We assume first that $\delta(\kappa)$ is greater than zero when κ equals zero, that is, capital depreciates even when it sits idle. Second, we assume that $\delta(\kappa)$ rises as κ increases above zero.[9] Thus, this curve shows that a higher κ leads to a higher depreciation rate, $\delta(\kappa)$.

The rate of return from owning capital, given by equation (8.6), equals the vertical distance between the green line and the blue curve in Figure 8.6. Owners of capital (households) select the utilization rate, κ, that maximizes this distance. In the graph, this maximization occurs at κ^*, where the vertical distance between the line and the curve is shown by the red arrows. Typically, κ^* will be set below its maximum feasible value of 8736 hours per year. The reason is that that this extremely high rate of utilization would result in rapid depreciation of the capital stock. The nearby box discusses other reasons for the utilization rate to be less than 100%.

Multiple shifts and overtime hours

We have argued that κ, the capital utilization, would typically be set at less than its maximum feasible value of 8736 hours per year. We got this result by considering the positive effect of κ on the depreciation rate, $\delta(\kappa)$. Other reasons for less than full utilization are also important. To consider these reasons, it is easiest to think about why a business that owns capital—factories and machines—would choose to operate this capital only part of the time.

We have assumed that real GDP, Y, depends on capital services in the form κK:

[9] We also assume that $\delta(\kappa)$ gets more sensitive to κ as κ increases. Graphically, the curve $\delta(\kappa)$ has a convex shape—it bows out toward the horizontal axis.

$$(8.3) \qquad Y = F(\kappa K, L).$$

If K represents the number of machines, then κ represents the hours per year that each machine operates. If we start with a κ of 2080 hours per year—where businesses use their capital 8 hours per day on weekdays—κ could be raised by operating more than one shift per day or by opening on weekends. However, more hours of operation per week incur additional costs, for example, the electric power needed to keep the lights on. We should subtract these "user costs" of capital from the expression for real profit in equation (8.4) to get

$$\Pi/P = A\cdot F[(\kappa K)^d, L^d)] - (w/P)\cdot L^d - (R/P)\cdot(\kappa K)^d - \textit{user costs of capital.}$$

The presence of the user costs—which increase with κ—can explain why capital would be operated less than full-time. That is, a business may prefer to have 100 machines operated half the time, rather than 50 machines operated full-time.

Another consideration is that, to raise κ, businesses typically have to operate their machines and factories at less convenient times—for example, during evenings or weekends. Typically, these times of operation are more expensive than standard business hours because workers have to be paid higher wage rates for night shifts or overtime hours. In addition, complementary services from other businesses, such as suppliers and transporters, may be unavailable at these times. (Lower electricity rates and less highway congestion at off-hours are offsetting factors.) If we allow for the high costs of operation at unusual hours, we get another reason why businesses choose to use their capital at less than maximal capacity. Machines and factories often sit idle at 4am because it is not worthwhile to pay the high wages needed to attract workers at 4am.

Now we have to figure out how a change in the real rental price, R/P, changes the choice of the capital utilization rate, κ^*. Suppose that R/P rises from $(R/P)_1$ to $(R/P)_2$. At $(R/P)_1$, the real rental payments per unit of capital, $(R/P)_1 \cdot \kappa$, are given by the green line from the origin in Figure 8.7. At the higher real rental price, $(R/P)_2$, the real rental payments per unit of capital, $(R/P)_2 \cdot \kappa$, are given by the brown line, which is steeper than the green line. The depreciation curve, $\delta(\kappa)$, is the blue curve, the same as in Figure 8.6. This curve does not shift when R/P changes.

When the real rental price is $(R/P)_1$, households maximize the vertical distance between the green line and the blue curve by picking the utilization rate κ_1^*. When the real rental price rises to $(R/P)_2$, households maximize by choosing the higher utilization rate κ_2^*. Thus, an increase in R/P raises the capital utilization rate, κ. The reason is that an upward shift in the real rental payment per unit of capital, $(R/P) \cdot \kappa$, makes it worthwhile to raise κ even though this greater utilization leads to a higher depreciation rate, $\delta(\kappa)$.

C. Market clearing and capital utilization

We considered before the effect of an increase in the technology level, A, on real GDP, labor input, and other variables. Now we can also consider the effect of an increase in A on the capital utilization rate, κ, and, hence, on the quantity of capital services, κK.

Figure 8.8 puts together our analyses of the demand for and supply of capital services. The vertical axis has the real rental price, R/P, and the horizontal axis shows the market demand for and supply of capital services. The two downward-sloping demand curves come from Figure 8.5. The blue curve corresponds to the technology level A_1, and the red curve corresponds to the higher technology level A_2. Note that the market demand for capital services, $(\kappa K)^d$, is higher at any R/P when A is higher.

The upward-sloping supply curve comes from Figure 8.7. The supply curve slopes up because an increase in the real rental price, R/P, motivates a higher capital utilization rate, κ. Hence, for a given stock of capital, K, an increase in R/P raises the quantity of capital services supplied, $(\kappa K)^s$.[10] Figure 8.7 also tells us that, for a given R/P, the choice of the utilization rate, κ, does not depend on the technology level, A. Therefore, an increase in A from A_1 to A_2 does not shift the supply curve for capital services in Figure 8.8.

When the technology level is A_1, Figure 8.8 shows that the market for capital services clears when the real rental price is $(R/P)_1$ and the quantity of capital services

[10] Notice a difference between labor services and capital services. An increase in the real wage rate, w/P, raised labor supply, L^s, because of the substitution effect between consumption and leisure. However, the higher w/P also had an income effect, which reduced L^s. Thus, the overall effect was ambiguous—an increase in w/P raised L^s if the income effect was not too strong. In contrast, there is no income effect on the capital utilization rate, κ, and, hence, on the supply of capital services, K^s. The reason is that, unlike workers, capital goods are inanimate objects that do not gain utility by working less. Therefore, an increase in the real rental price, R/P, unambiguously raises κ and, thereby, the quantity of capital services supplied, K^s.

is $(\kappa K)_1$. When the technology level rises to A_2, the demand curve for capital services shifts to the right, and the supply curve does not shift. Therefore, the market clears at the higher real rental price, $(R/P)_2$, and the larger quantity of capital services, $(\kappa K)_2$.[11]

We noted in chapter 7 that an increase in the technology level raised the real rental price, *R/P*. The new effect is the increase in the quantity of capital services. Since the capital stock, *K*, is fixed in the short run, the increase in capital services reflects the rise in the utilization rate—from κ_1^* to κ_2^* in Figure 8.7. We therefore have that booms—where a high *A* causes real GDP to be high—will have a relatively high capital utilization rate, whereas recessions—where a low *A* causes real GDP to be low—will have a relatively low utilization rate.

Recall that the production function is

$$(8.3) \qquad Y = A \cdot F(\kappa K, L).$$

We now have three reasons why real GDP rises in a boom and falls in a recession. First, a high or low technology level, *A*, causes real GDP to be correspondingly high or low, for given inputs of capital services, κK, and labor, *L*. Second, a high or low *A* causes *L* to be correspondingly high or low. These procyclical changes in labor input help the economy to produce more goods in a boom and fewer goods in a recession. Third, a high or low *A* raises the capital utilization rate, κ, and, thereby, the quantity of capital services, κK. These procyclical changes in κK also help the economy to produce more goods in a boom and fewer goods in a recession.

---

[11] The rise in capital services, κK, tends to increase the marginal product of labor, MPL. This rise in the MPL (at a given *L*) causes further movements to the right of the labor-demand curve in Figure 8.3. Hence, the quantity of labor, *L*, and the real wage rate, *w/P*, rise by more than before. Similarly, the increase in labor input, *L*, tends to raise the marginal product of capital services, MPK. The rise in the MPK (at a given κK) causes further movements to the right of the demand curve for capital services in Figure 8.7. Hence, the quantity of capital services, κK, and the real rental price of capital, *R/P*, rise by more than before.

D. The cyclical behavior of capacity utilization

To check our prediction that the capital utilization rate, κ, is procyclical, we can use the Federal Reserve's data on capacity utilization rates in manufacturing, mining, and public utilities.[12] These estimates measure the percentage of "normal capacity" at which businesses are operating. From January 1948 to September 2002, the average of the overall capacity utilization rate was 82%, with a range from 71% to 92%.

The blue graph in Figure 8.9 shows the deviation of the log of the capacity utilization rate from its trend (using our usual method to construct the trend). The red graph is again the deviation of the log of real GDP from its trend. Note that the capacity utilization rate is clearly procyclical.[13] Thus, the model's prediction about the cyclical behavior of the capital utilization rate matches up with the data on capacity utilization rates.[14]

III. Summing Up

We extended the model from chapter 7 to allow for variable labor supply and variable utilization of capital. With variable labor supply, the model can match the observation that employment and worker-hours are procyclical. However, these results

[12] The Federal Reserve computes capacity utilization by comparing a sector's current output of goods with an estimate of the capacity of each sector to produce goods. The information on output comes from the Federal Reserve's data on industrial production.

[13] From 1954.1 to 2004.3, the correlation of the cyclical part of the capacity utilization rate with the cyclical part of real GDP is 0.91. The capacity utilization rate is more variable than real GDP—the standard deviation of the cyclical part is 3.3%, compared to 1.6% for the cyclical part of real GDP.

[14] We should note some problems with the data on capacity utilization rates. First, the Federal Reserve measures utilization only in manufacturing, mining, and public utilities. Since the first two of these sectors are more volatile than the overall economy, the variability of the measured capacity utilization rate probably overstates the variability of the capacity utilization rate for the whole economy. Second, the data on capacity utilization are based partly on measures of outputs by sector. Thus, there is a sense in which the method of measurement forces the constructed capacity utilization rate to move along with output and, hence, real GDP.

depend on a positive response of labor supply to the real wage rate. This response applies only if the substitution effects from changes in the real wage rate dominate the income effects. With variable capital utilization, the model can match the finding that capacity utilization rates are procyclical. This result applies because the procyclical behavior of the real rental price generates a procyclical pattern in capital utilization and, hence, in the quantity of capital services supplied. The cyclical variations in inputs of labor and capital services contribute to the rise of real GDP in a boom and the decline of real GDP in a recession.

Questions and Problems

Mainly for review

8.1 Discuss the effects on this year's labor supply, L_1^s, from the following changes:

a. An increase in the interest rate, i_1.

b. A permanent increase in the real wage rate, (w/P).

c. An increase in the current real wage rate, $(w/P)_1$, but no change in future real wage rates.

d. A one-time windfall, which raises initial assets, $(B_0/P + K_0)$.

8.2. Explain how the quantity of capital services depends on the stock of capital, K, and the capital utilization rate, κ. Why is the rate of return on capital given by the expression shown in equation (8.6)?

8.3. Use Figure 8.6 to study the capital utilization rate, κ. How does κ change when

a. the real rental price, R/P, rises?

b. the depreciation rate, $\delta(\kappa)$, rises for each value of κ?

Problems for discussion

8.x. Labor-force participation

For the United States in the post-World War II period, a major part of the increase in labor input, *L*, reflected a rise in labor-force participation, particularly by women. The table shows how the overall participation rate changed from 1950 to 2000. (The participation rate is defined as the ratio of the civilian labor force to the non-institutional population, which is the population 16 and over not residing in prisons and other institutions.)

Year	**Participation rate (%)**
1950	59.2
1960	59.4
1970	60.4
1980	63.8
1990	66.5
2000	67.1

a. Is this behavior of labor-force participation consistent with our analysis of income and substitution effects?

b. Is the behavior consistent with the observation that average weekly hours of workers changed little in the United States in the post-World War II period (as noted in appendix B of this chapter)?

c. Are there other factors that we should bring in to explain the changing behavior of labor-force participation?

8.x. A change in population

This question is an extended version of problem 7.x. Assume a one-time decrease in population, possibly caused by an onset of plague or a sudden out-migration.

a. Use a variant of Figure 8.3 to determine the effects on the labor market. What happens to labor input, L, and the real wage rate, w/P?

b. Use a variant of Figure 8.8 to determine the effects on the market for capital services. What happens to capital input, κK, and the real rental price, R/P? What happens to the interest rate, i?

c. What happens to output, Y, and consumption, C? What happens to investment, I? What happens over time to the stock of capital, K?

8.x. A change in the capital stock

This question is an extended version of problem 7.x. Assume a one-time decrease in the capital stock, possibly caused by a natural disaster or an act of war. Assume that population does not change.

a. Use a variant of Figure 8.8 to determine the effects on the market for capital services. What happens to capital input, κK, and the real rental price, R/P? What happens to the interest rate, i?

b. Use a variant of Figure 8.3 to determine the effects on the labor market. What happens to labor input, L, and the real wage rate, w/P?

c. What happens to output, Y, and consumption, C? What happens to investment, I? What happens over time to the stock of capital, K?

8.x. A shift in desired saving

This question is an extended version of problem 7.x. Suppose that households change their preferences so that they wish to consume more and save less in the current year. That is, current consumption, C_1, rises for a given interest rate and for given current and future income.

a. Use a variant of Figure 8.3 to determine the effects on the labor market. What happens to labor input, L, and the real wage rate, w/P?

b. Use a variant of Figure 8.8 to determine the effects on the market for capital services. What happens to capital input, κK, and the real rental price, R/P? What happens to the interest rate, i?

c. What happens to consumption, C, and investment, I? What happens over time to the stock of capital, K?

Appendix

A. The cyclical behavior of labor productivity

In this and the previous chapter, we described the behavior of a number of economic variables during economic fluctuations. Another macroeconomic variable that is often discussed is labor productivity. Usually, this concept refers to the average product of labor, Y/L. We consider here what the model predicts for the behavior of Y/L during booms and recessions. Then we compare the predictions with the data.

Consider again the economy's response to an increase in the technology level, A. Recall that real GDP is given by

(8.3) $$Y = A \cdot F(\kappa K, L).$$

We found that an increase in A raises real GDP, Y, partly because of the direct effect from the higher A and partly because of the increases in the two inputs, capital services, κK, and labor, L. For given L, increases in A and κK raise Y and, thereby, increase the average product of labor, Y/L. However, for given A and κK, increases in L tend to reduce Y/L. That is, usual production functions exhibit diminishing average product of labor—higher L leads to lower Y/L. Overall, we predict that a rise in A will increase the average product of labor, Y/L, unless the negative effect from higher L dominates the other influences.

We can go further to resolve this ambiguity, because we also know that an increase in the technology level, A, raises the market-clearing real wage rate, w/P—see Figure 8.3. Moreover, w/P equals the marginal product of labor, MPL. Therefore, an increase in A must raise the MPL. For most forms of production functions, the average product of labor, Y/L, moves in the same direction as the marginal product, MPL. Therefore, we anticipate that the rise in A will raise Y/L.

We can get an example of the relation between the average and marginal products of labor by considering the Cobb-Douglas production function, which we discussed in the appendix to chapter 3:

$$Y = AK^{\alpha}L^{1-\alpha}, \tag{3.17}$$

where $0<\alpha<1$. We can apply this production function to our present analysis if we replace the capital stock, K, by the quantity of capital services, κK:

$$Y = A\cdot(\kappa K)^{\alpha}\cdot L^{1-\alpha}. \tag{8.6}$$

The average product of labor can be found by dividing each side of equation (8.6) by L:

$$Y/L = A\cdot(\kappa K)^{\alpha}\cdot L^{-\alpha}$$

$$Y/L = A\cdot(\kappa K/L)^{\alpha}. \tag{8.7}$$

The marginal product of labor, MPL, can be found by differentiating the right-hand side of equation (8.6) with respect to L, while holding fixed A and κK:

$$MPL = dY/dL = (1\text{-}\alpha)\cdot A\cdot(\kappa K)^{\alpha}\cdot L^{-\alpha}$$

$$MPL = (1\text{-}\alpha)\cdot A\cdot(\kappa K/L)^{\alpha}. \tag{8.8}$$

A comparison of equation (8.8) with equation (8.7) shows that the marginal product of labor, MPL, is the constant multiple $1-\alpha$ of the average product, Y/L. Since we know that a rise in the technology level, A, leads to an increase in the MPL, we also know that the rise in A raises the average product of labor, Y/L. Hence, with a Cobb-Douglas production function, the model predicts that a rise in A will increase Y/L. Thus, Y/L will be procyclical: high in booms and low in recessions. This conclusion tends to hold also for other forms of production functions.

To compare this prediction with the data, we can use two measures of the average product of labor, *Y/L*. These measures correspond to the two concepts of labor input, *L*, that we used before. The first measure of *L* is employment (numbers of persons with jobs), so that *Y/L* is real GDP per worker. The second measure of *L* is total hours worked (which multiplies employment by average weekly hours worked per employee). In this case, *Y/L* is real GDP per worker-hour.

As before, we use the method from Figure 7.2 to detrend the log of each variable. Then we calculate the deviations of the log of each variable from its trend to get measures of cyclical parts. We show the cyclical part of real GDP per worker as the blue graph in Figure 8.10. The red graph shows the cyclical part of real GDP. We see that real GDP per worker is procyclical. That is, as our theory predicts, this measure of the average product of labor fluctuates in the same direction as real GDP.[15]

We show the cyclical part of real GDP per worker-hour as the blue graph in Figure 8.11. The cyclical part of real GDP is again given by the red graph. In this case, real GDP per worker-hour is procyclical, but the correlation with the cyclical part of real GDP is not that strong. That is, the movements in this measure of the average product of labor do not relate closely to the fluctuations in real GDP.[16] Thus, real GDP per worker-hour may not be as procyclical as our model predicts.

[15] From 1954.1 to 2004.3, the correlation of the cyclical part of real GDP per worker with the cyclical part of real GDP is 0.57. The standard deviation of the cyclical part of real GDP per worker is 1.0%, compared with 1.6% for real GDP.
[16] From 1964.1 to 2004.3 (where we have data on average hours worked), the correlation of the cyclical part of real GDP per worker-hour with the cyclical part of real GDP is 0.31. The standard deviation of the cyclical part of real GDP per worker-hour is 0.7%.

B. The interest rate, the real wage rate, the saving rate, and the quantity of labor in the Solow growth model

In chapters 3-5, we used the Solow model to understand how economies grew in the long run. Now we can use our analyses of microeconomic foundations and market clearing to gain additional insights about the path of economic growth. In particular, we can study the transitional behavior of the interest rate, the real wage rate, the saving rate, and the quantity of labor input. We begin by neglecting the responsiveness of labor supply, L^s, to substitution and income effects. In this case, total labor, L, rises over time due to population growth but otherwise does not change. We also neglect any responsiveness of the capital utilization rate, κ, to the real rental price, R/P.

1. The interest rate. To work out the transition path of the interest rate, i, we can use the version of the Solow model with population growth at the constant rate n but with no technological progress. In this case, the technology level, A, is fixed. Recall that the transition to the steady state features increases in real GDP and capital per worker, $y = Y/L$ and $k = K/L$. For fixed A, the rising path of k implies a declining path of the marginal product of capital, MPK. (See Figure 3.7 in chapter 3.) Figure 8.12 illustrates this effect by considering a rise in capital per worker from k_1 to k_2. The rise in k lowers the marginal product of capital, MPK. Since the MPK equals the market-clearing real rental price, R/P, the figure shows that R/P declines from $(R/P)_1$ to $(R/P)_2$.

Recall that the interest rate, i, equals the rate of return from owning capital, which is given by

(8.6) $$\textit{rate of return from owning capital} = (R/P)\cdot\kappa - \delta(\kappa).$$

Since κ is assumed to be fixed, the declining path of the real rental price, R/P, leads to a declining path of i. Hence, we have a new prediction from the Solow model: the interest rate falls during the transition to the steady state.

2. The real wage rate. An increase in capital per worker, k, lowers the marginal product of capital, MPK, because of the diminishing marginal product of capital. However, as k rises, each worker has more capital to work with. Therefore, the marginal product of labor, MPL, tends to increase. Figure 8.13 shows accordingly that a rise in k from k_1 to k_2 leads to an increase in the MPL. Since the MPL equals the market-clearing real wage rate, w/P, the figure shows that w/P rises from $(w/P)_1$ to $(w/P)_2$. Thus, we have another prediction from the Solow model: the real wage rate rises during the transition to the steady state.

3. The saving rate. The Solow model assumes that the gross saving rate, s, is constant. Now we can use the analysis from chapter 7 to see how households change their saving rates during the transition to the steady state.[17] We can carry out this analysis in terms of the two forces that we have emphasized—the intertemporal-substitution effect and the income effect.

We found that increases in capital per worker, k, lead to decreases in the interest rate, i. We know that a lower i means a weaker intertemporal-substitution effect. That is, as i declines, households have less incentive to defer consumption from the present to the

[17] The extension to allow a variable saving rate was the major contribution of the neoclassical growth model developed in the mid 1960s. Two major contributors were David Cass and Tjalling Koopmans.

future. Therefore, households tend to consume more in relation to their current income. For this reason, the gross saving rate, s, tends to fall.

Our analysis is incomplete because we have neglected an income effect. To understand this effect, think about the initial capital per worker, $k(0)$. If $k(0)$ is far below its steady-state value, k^*, real GDP per worker, $y(0) = A \cdot f[k(0)]$, is well below its steady-state level, $y^* = A \cdot f(k^*)$. If the relation between workers and population—that is, the labor-force participation rate—is fixed, the low real GDP per worker translates into a low real GDP per person. Since real GDP per person determines real income per person, we have that real income per person is initially far below its steady-state level. Households consume in relation to long-run income, which is an average of prospective incomes over all years.[18] Since steady-state income is well above current income, long-run income will also be well above current income. Therefore, households will choose a level of current consumption that is high compared to current income. Hence, today's gross saving rate, s, tends to be low.

As k rises toward k^*, real GDP per worker, $y = A \cdot f(k)$, gets closer to its steady-state level, y^*. Therefore, current income per person gets closer to long-run income per person. Since households choose consumption in relation to long-run income, the ratio of consumption to current income falls. Hence, the gross saving rate, s, rises.

The overall effect of increasing capital per worker, k, on the gross saving rate, s, depends on the balance of the intertemporal-substitution effect and the income effect. The declining interest rate suggests that the gross saving rate, s, would fall as k increases, whereas the rise in real GDP per worker, y, toward its steady-state level, y^*, suggests that

[18] We mentioned in chapter 7 that Milton Friedman referred to this long-run income as permanent income.

s would rise as *k* increases. We cannot be sure, on theoretical grounds, whether the net effect is for the gross saving rate to rise or fall as *k* increases.

Let's see whether we can resolve the ambiguity by looking at international data. Figure 8.14 shows the relation between a country's gross saving rate—measured by the average ratio of gross investment to GDP from 1985 to 1994—and the level of real GDP per person in 1990.[19] We see that a higher real GDP per person matches up with a higher investment ratio. However, this pattern could result because a higher saving rate raises real GDP per person, rather than the reverse. In fact, the Solow model predicts that a country that saves at a higher rate would have a higher real GDP per person in the steady state.

A more detailed analysis of the data shows that an increase in real GDP per person does lead to an increase in the saving rate. However, this effect applies only over a range of real GDP per person. For the richest countries, the effect is close to zero.

We can modify the Solow model to take account of the empirical relation between the gross saving rate, *s*, and capital per worker, *k*. To reflect our reading of the cross-country evidence, assume that *s* is given by an upward-sloping function, *s(k)*. Assume, however, that the function *s(k)* becomes flat at high values of *k*. With this amendment, the basic equation for the Solow model, from equation (4.8) of chapter 4, becomes

$$\Delta k/k = s(k)\cdot(y/k) - (\delta+n). \qquad (8.9)$$

The new feature is that we replaced the constant saving rate, *s*, by the saving function, *s(k)*.

[19] In a closed economy, the saving rate equals the investment ratio. In an open economy, national saving can depart from domestic investment. However, for most countries and time periods, the deviations are not large when considered in relation to GDP. Therefore, we get similar empirical results if we replace the investment ratio by the national saving rate.

We know from our discussion of the Solow model in chapter 3 that a rise in capital per worker, k, reduces the average product of capital, y/k. This effect tends to lower the growth rate, $\Delta k/k$, in equation (8.9). However, this effect is now offset by a rise in the gross saving rate, given by $s(k)$. Therefore, a tendency for the gross saving rate to rise with k slows down the process of convergence. Or, to put it another way, a rising gross saving rate is one reason why richer societies can sustain economic growth for a long time. However, this effect does not go on forever, because $s(k)$ becomes flat when k attains a very high value. Therefore, we still predict that, in the absence of technological progress, the growth rate of k would fall eventually toward zero.

4. Labor input. We assumed in our analysis of the Solow model that each person worked a fixed amount of time. Now we can use the analysis from the current chapter to see how work effort varies as an economy develops. As usual, the responses can be analyzed in terms of substitution and income effects. The important substitution effect is that the increases in the real wage rate, w/P, motivates higher labor supply. The income effect is that increases in real GDP per worker, y, motivate lower labor supply. In general, the overall effect is ambiguous. That is, we cannot be sure on theoretical grounds whether people will work more or less as an economy develops.

Empirically, average hours worked per worker tend to decline during early stages of economic development. In the United States, average weekly work hours for production workers in manufacturing decreased from 55 to 60 in 1890 to about 50 in

1914, 44 in 1929, and 38 in 1940. Since World War II, the numbers leveled off. Average weekly hours ranged between 39 and 42, and the value in 2004 was 41.[20]

A roughly similar pattern applies to the United Kingdom. Average weekly hours of male manual workers in manufacturing went from 60 in 1850 to 55 in 1890, 54 in 1910, 48 in 1938, and 43 in 2002.[21] Again, there was a long-term decline and a possible leveling off in the long run.

In contrast to the United States and the United Kingdom, some other rich countries—notably, Germany, France, and Italy—continued to have declines in average hours worked per person in recent years. Prescott (2004, Tables 3 and 4) discusses these patterns and attributes them to differences in tax policies, notably to rising tax rates in continental Western Europe.

We can also get information by comparing a group of rich countries with a group of poor countries at a point in time. Winston (1966, Table 1) found that, in 1953-60, the mean for ten rich countries for average weekly hours worked in manufacturing was 43.9. (The ten are the United States, Canada, Switzerland, Sweden, New Zealand, the United Kingdom, Norway, France, West Germany, and the Netherlands.) In the same period, the average for ten less-developed countries was 47.4. (These ten are Yugoslavia, Colombia, the Philippines, El Salvador, Ecuador, Guatemala, Peru, Taiwan, Egypt, and Ceylon.) Thus, the richer countries had lower average work hours.

---

[20] The data are from U.S. Department of Commerce (1975, pp. 168, 169). Values since 1939 are from the Bureau of Labor Statistics. The data available over the long term apply only to production workers in manufacturing. Average weekly hours in the total private economy are lower, 34 in 2004.

[21] The earlier data are rough averages from Bienefeld (1972, chapters 4 and 5). The figure for 1938 is from Mitchell and Jones (1971, p. 148). Recent values are from Central Statistical Office, *Annual Abstract of Statistics*. The data available over the long term apply only to male manual workers in manufacturing. In 2003, average weekly hours in all industries were 40.

From the perspective of the model, we infer that the income effect from increasing real GDP per person tends to dominate over the substitution effect from a rising real wage rate at least through middle stages of economic development. Hence, until a country becomes rich, average hours worked per person tend to decrease as a country develops. However, when a country becomes rich enough—roughly the level for the United States around World War II—the substitution effect from a rising real wage rate roughly balances the income effect. Thus, in this range, average hours worked per person change little as a country gets richer. However, if a country increasingly discourages work effort—as in the case of rising tax rates in continental Western Europe—the decline in average hours worked per person can continue even when a country is rich.

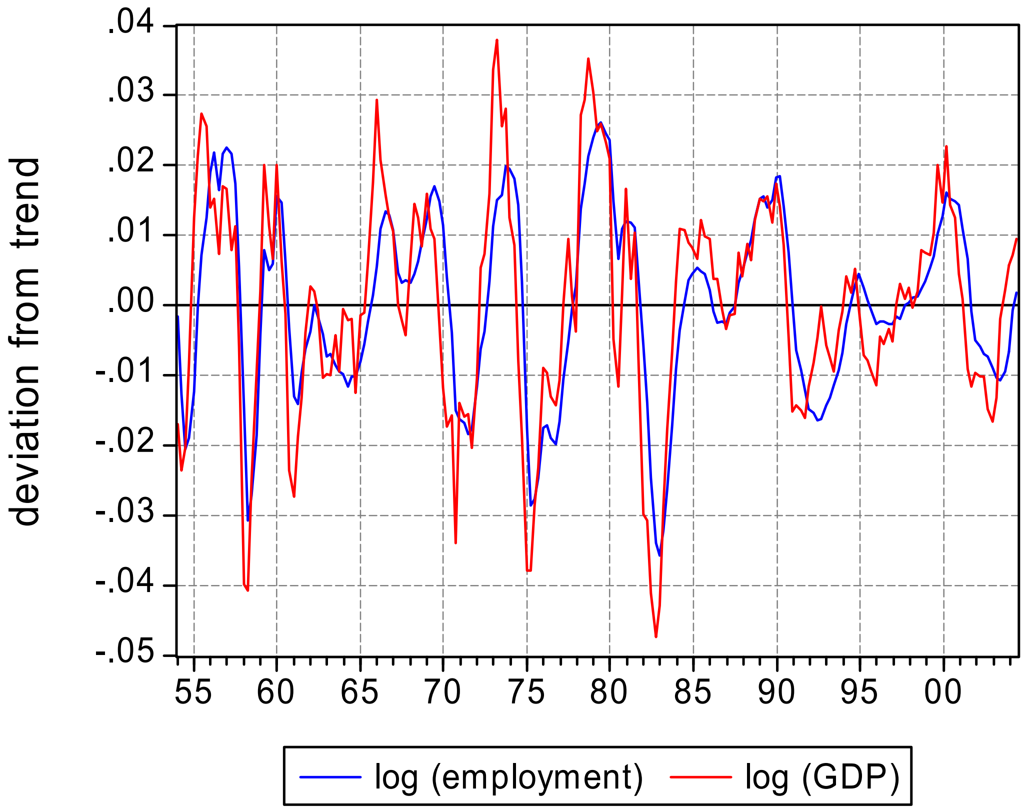


Figure 8.1

Cyclical Parts of U.S. GDP and Employment

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of employment from its trend. Employment comes from the BLS payroll survey—it measures the total number of persons working on payrolls in the non-agricultural economy. The data on real GDP and employment are quarterly and seasonally adjusted. (The underlying data on employment are monthly.) Employment is procyclical—it fluctuates closely with real GDP and is nearly as variable as real GDP.

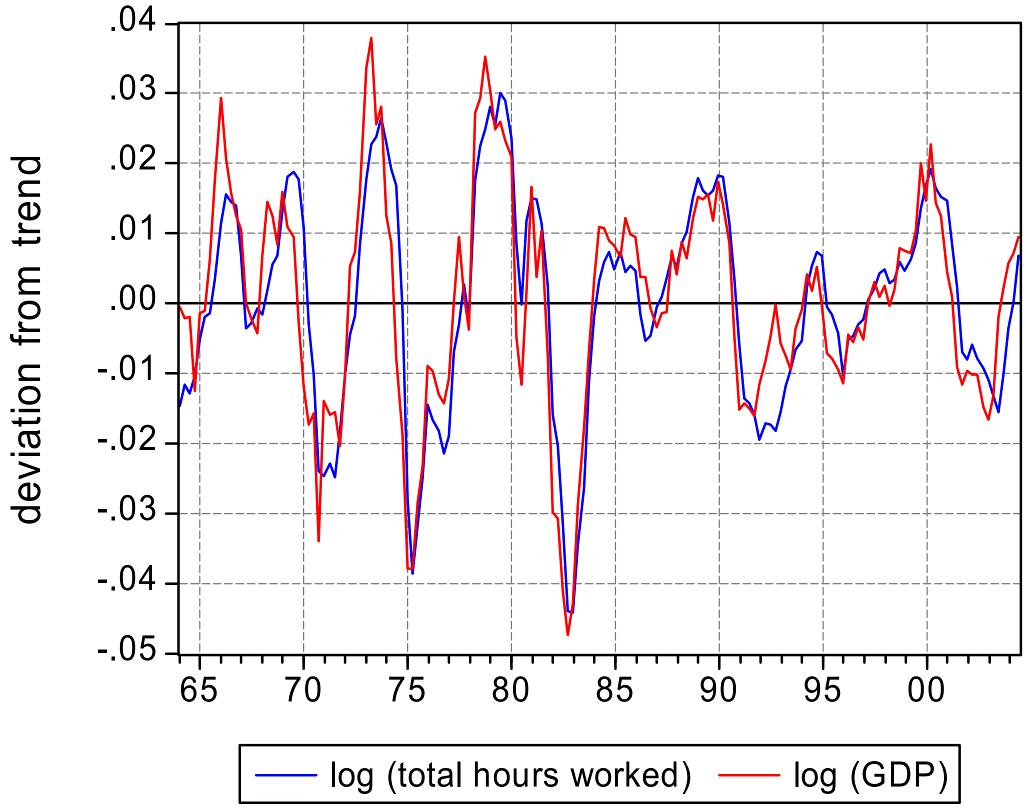


Figure 8.2

Cyclical Parts of U.S. GDP and Total Hours Worked

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of total hours worked from its trend. Total hours worked is employment (the concept in Figure 8.1) multiplied by average weekly hours of persons working. The data for weekly hours come from the BLS payroll survey and refer to the private, non-agricultural economy. These numbers are available since 1964. The data on real GDP and total hours are quarterly and seasonally adjusted. (The underlying data on total hours are monthly.) Total hours worked is procyclical—it fluctuates closely with real GDP and is about as variable as real GDP.

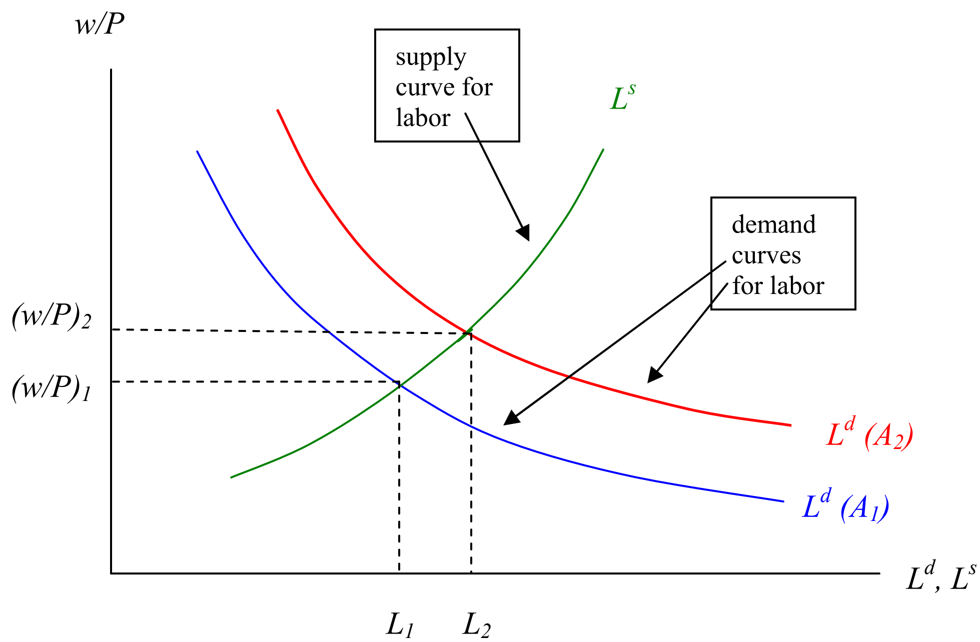


Figure 8.3

Clearing of the Labor Market

At the technology level A_1, the demand for labor, labeled $L^d\ (A_1)$ along the blue curve, slopes downward versus the real wage rate, w/P. At the higher technology level, A_2, the demand for labor, labeled $L^d\ (A_2)$ along the red curve, is larger at any given w/P. These two curves are the same as those shown in Figure 7.5. The supply of labor, L^s, shown in green, slopes upward versus w/P because we assume that the substitution effect dominates over the income effect. Thus, the increase in the technology level from A_1 to A_2 raises the real wage rate from $(w/P)_1$ to $(w/P)_2$ and raises labor input from L_1 to L_2.

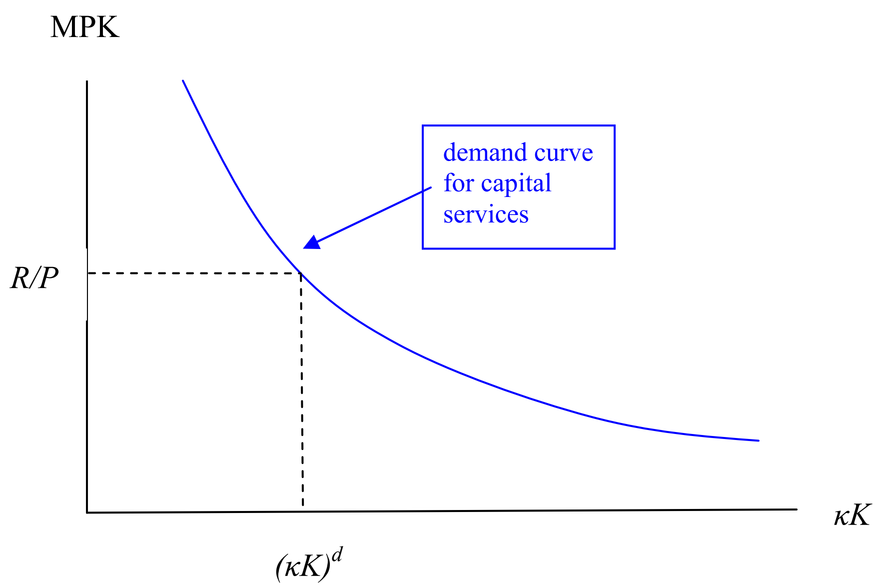


Figure 8.4

Demand for Capital Services

For a given technology level, A, and labor input, **L**, the marginal product of capital services, MPK, decreases as the quantity of capital services, κK, rises. The household chooses the quantity of capital services, $(\kappa K)^d$, where the MPK equals the real rental price, R/P.

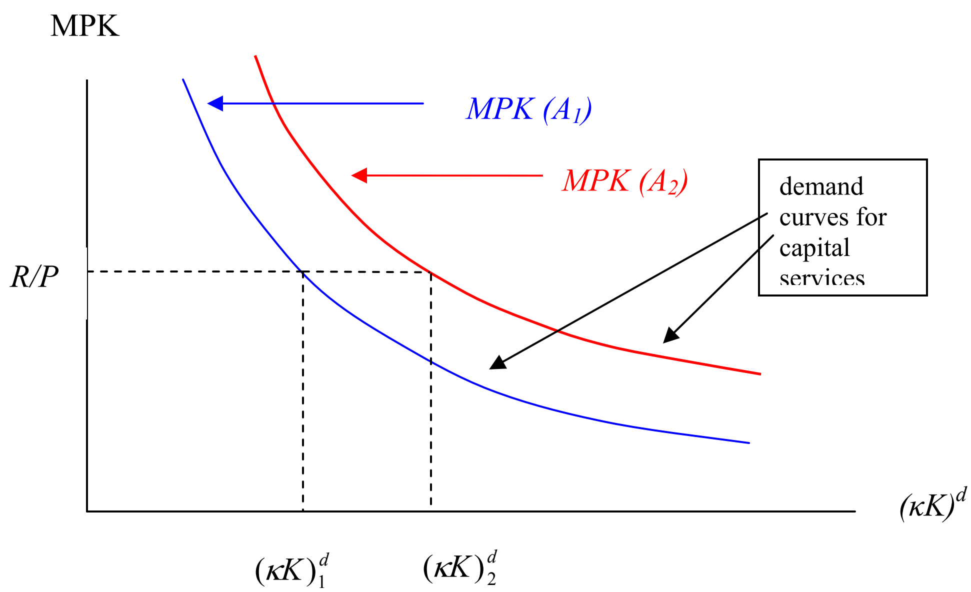


Figure 8.5

Effect of a Shift in the Technology Level on the Demand for Capital Services

When the technology level is A_1, the marginal product of capital services, MPK, is given by the blue curve, labeled as *MPK* (A_1). This curve is the same as the one in Figure 8.4. At the real rental price *R/P*, shown on the vertical axis, the quantity of capital services demanded is $(\kappa K)_1^d$ on the horizontal axis. The technology level A_2 is greater than A_1. Therefore, the MPK along the red curve labeled *MPK* (A_2) is higher at any level of capital services than the value along the blue curve. When the technology level is A_2 and the real rental price is *R/P*, the quantity of capital demanded is $(\kappa K)_2^d$, which is greater than $(\kappa K)_1^d$.

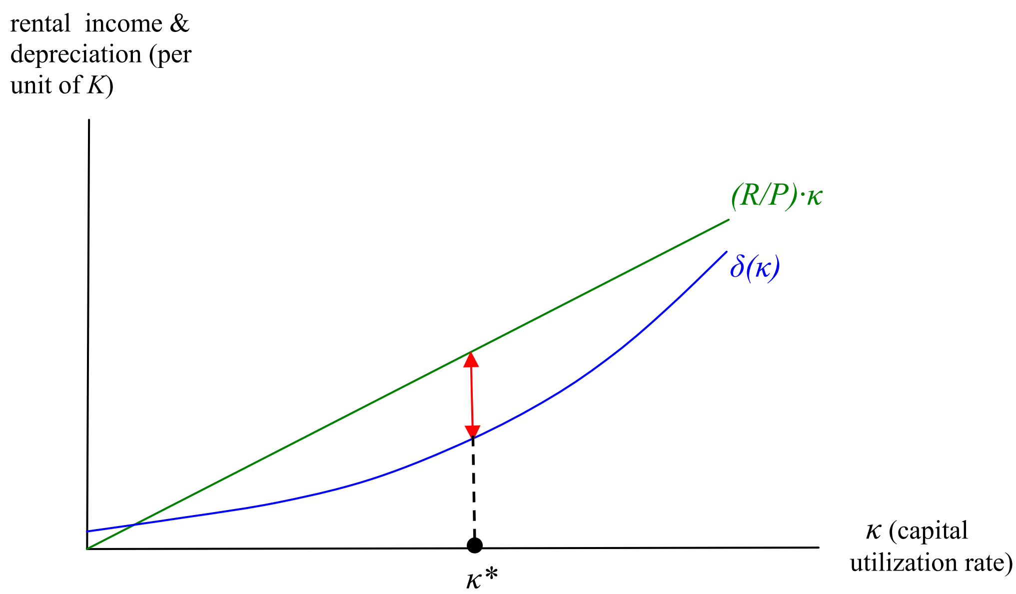


Figure 8.6

Choosing the Capital Utilization Rate

The green line from the origin is the real rental income per unit of capital, $(R/P)\cdot\kappa$. The blue curve shows the depreciation rate, $\delta(\kappa)$, as an upward-sloping function of the capital utilization rate, κ. Owners of capital choose κ to maximize the vertical distance between the line and the curve—this occurs at κ^*.

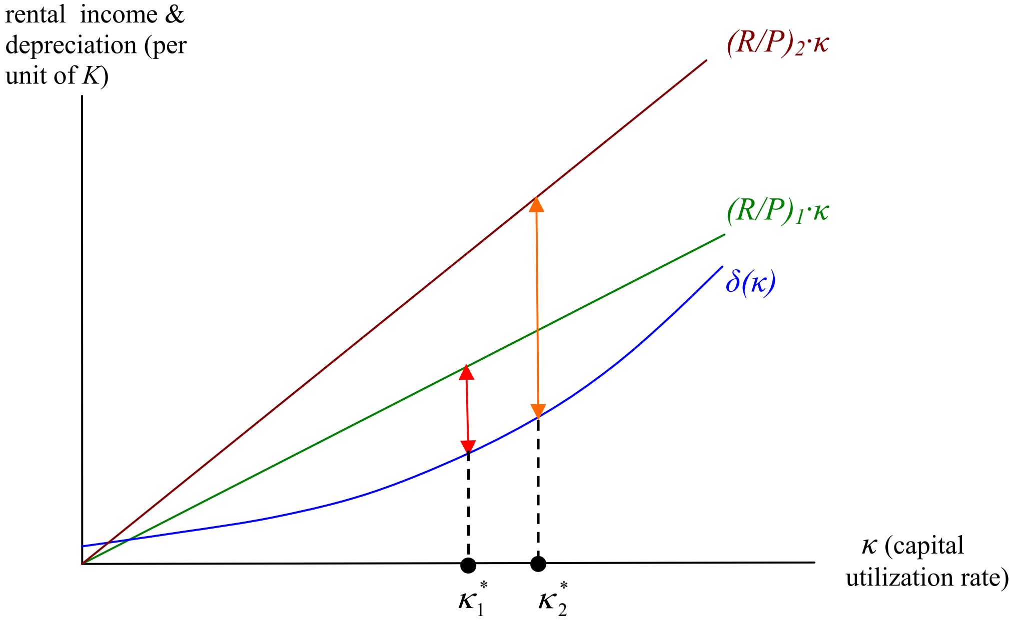


Figure 8.7

Effect of an Increase in the Real Rental Price on the Capital Utilization Rate

The blue curve is the depreciation rate, $\delta(\kappa)$, as graphed in Figure 8.6. The two lines from the origin are the real rental income per unit of capital, $(R/P)\cdot\kappa$. The green line is for the real rental price $(R/P)_1$, and the brown line is for the higher real rental price $(R/P)_2$. At $(R/P)_1$, owners of capital maximize the difference between the rental income line and the depreciation curve by choosing the capital utilization rate κ_1^* (see the red arrows). At $(R/P)_2$, they maximize the difference by choosing the higher utilization rate κ_2^* (see the orange arrows). Therefore, an increase in the real rental price, R/P, raises the capital utilization rate, κ.

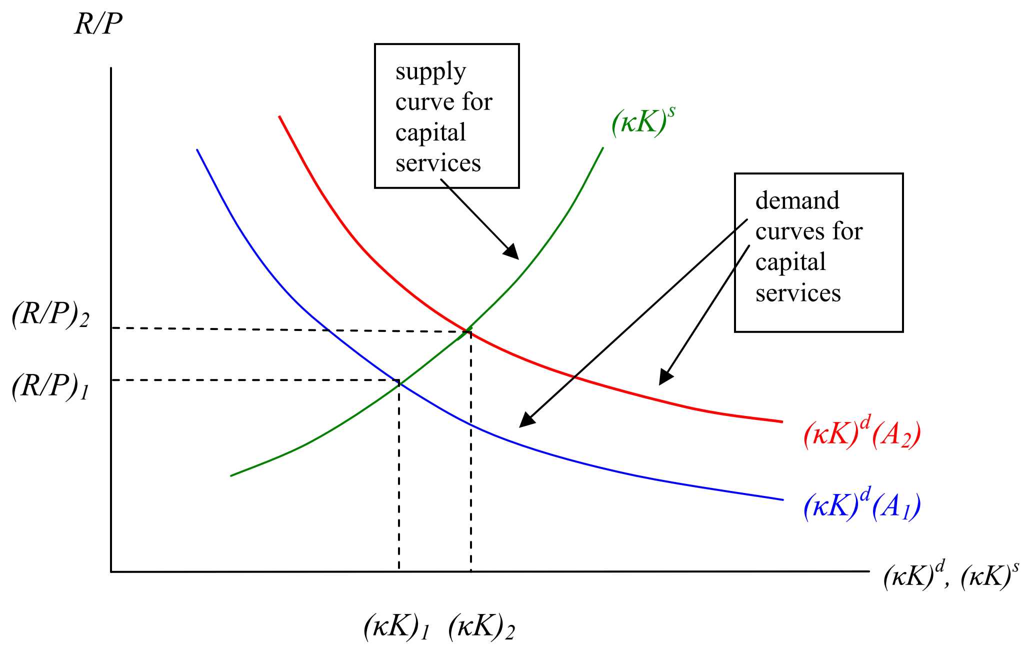


Figure 8.8

Clearing of the Market for Capital Services

At the technology level A_1, the demand for capital services, labeled $(\kappa K)^d(A_1)$ along the blue curve, slopes downward versus the real rental price, *R/P*. At the higher technology level A_2, the demand for capital services, labeled $(\kappa K)^d(A_2)$ along the red curve, is larger at any given *R/P*. These curves are the same as those in Figure 8.5. The supply of capital services, $(\kappa K)^s$, slopes upward versus *R/P* along the green curve because an increase in *R/P* raises the capital utilization rate, κ (see Figure 8.7). Thus, an increase in the technology level from A_1 to A_2 raises the market-clearing real rental price from $(R/P)_1$ to $(R/P)_2$ and the quantity of capital services from $(\kappa K)_1$ to $(\kappa K)_2$. Since the stock of capital, *K*, is fixed, the increase in capital services reflects the rise in the utilization rate from κ_1 to κ_2, as in Figure 8.7.

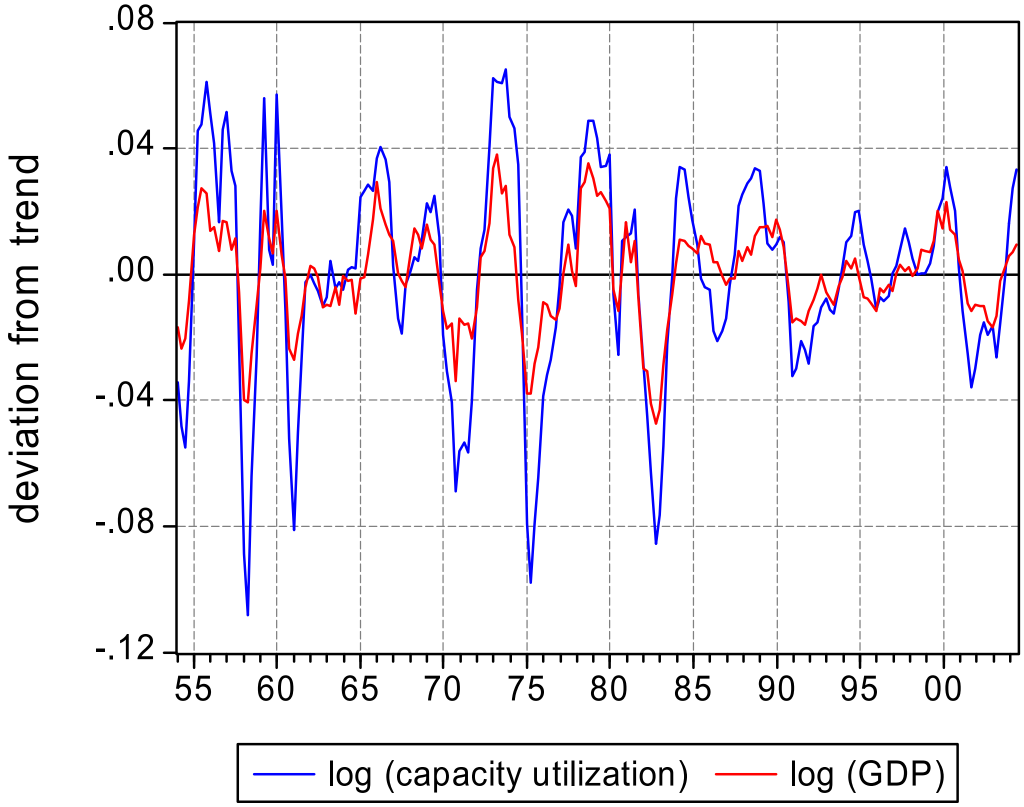


Figure 8.9

Cyclical Parts of U.S. GDP and Capacity Utilization

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of the capacity utilization rate from its trend. These data are based on industrial production and apply to manufacturing from 1948 to 1966 and to a broader index that includes mining and public utilities from 1967 to 2004. The data on real GDP and capacity utilization are quarterly and seasonally adjusted. (The underlying data on capacity utilization are monthly.) The capacity utilization rate is procyclical—it fluctuates closely with real GDP but is more variable than real GDP.

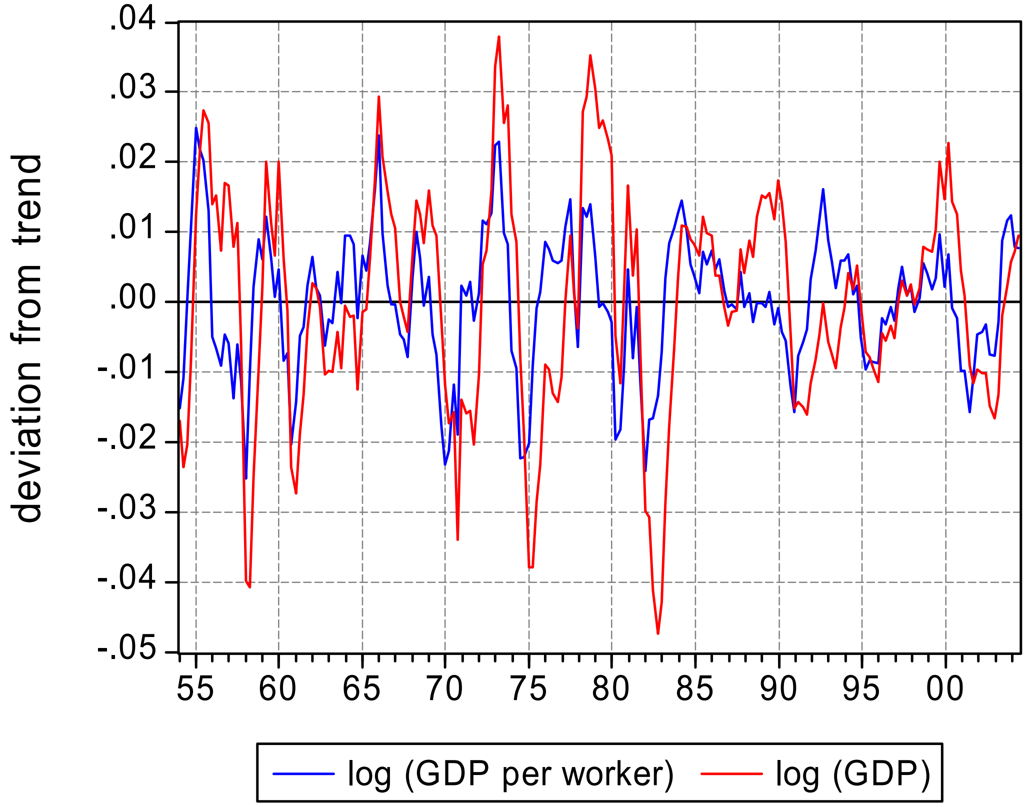


Figure 8.10

Cyclical Parts of U.S. GDP and GDP per Worker

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of real GDP per worker from its trend. (The concept for numbers of workers is the one used in Figure 8.1) Real GDP per worker is procyclical—it fluctuates with real GDP but is less variable than real GDP.

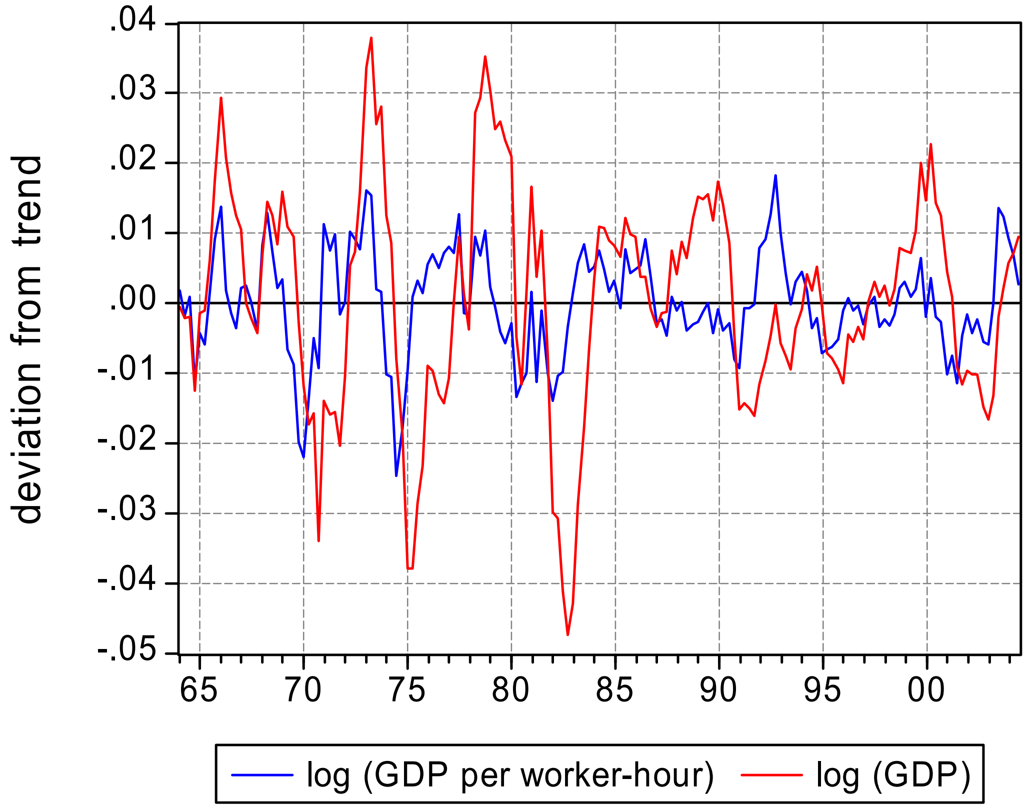


Figure 8.11

Cyclical Parts of U.S. GDP and GDP per Worker-Hour

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of real GDP per worker-hour from its trend. (The concept for worker-hours is the one used in Figure 8.2.) Real GDP per worker-hour is weakly procyclical—it fluctuates weakly with real GDP and is less variable than real GDP.

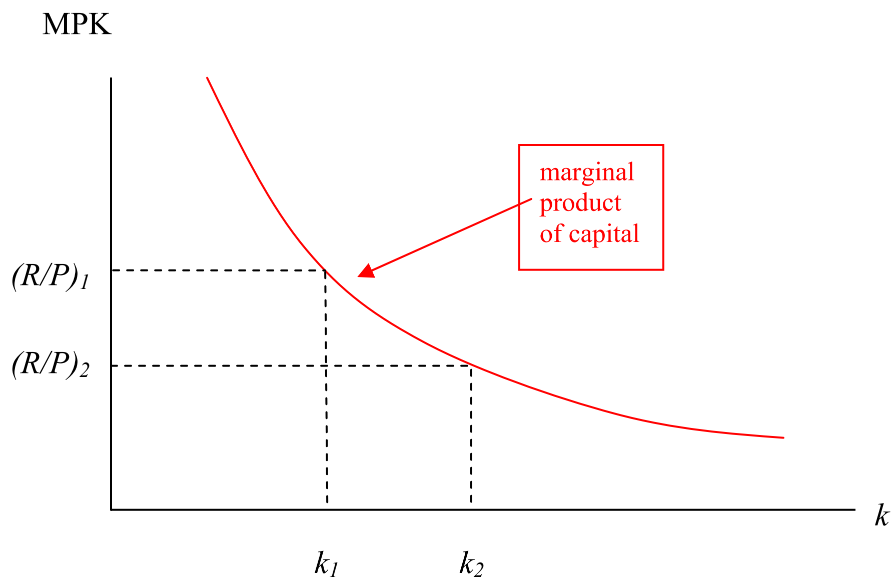


Figure 8.12

Effect of Higher Capital per Worker on the Real Rental Price

In the Solow growth model, the capital-labor ratio, k, rises over time, for example, from k_1 to k_2. The rise in k reduces the marginal product of capital, MPK, and thereby lowers the market-clearing real rental price from $(R/P)_1$ to $(R/P)_2$.

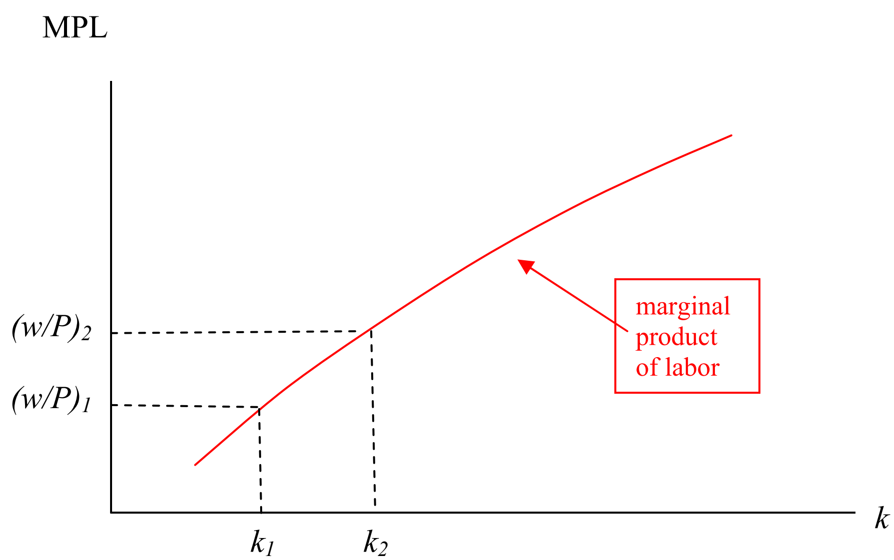


Figure 8.13

Effect of Higher Capital per Worker on the Real Wage Rate

In the Solow growth model, the capital-labor ratio, k, rises over time, for example, from k_1 to k_2. The rise in k raises the marginal product of labor, MPL, and thereby increases the market-clearing real wage rate from $(w/P)_1$ to $(w/P)_2$.

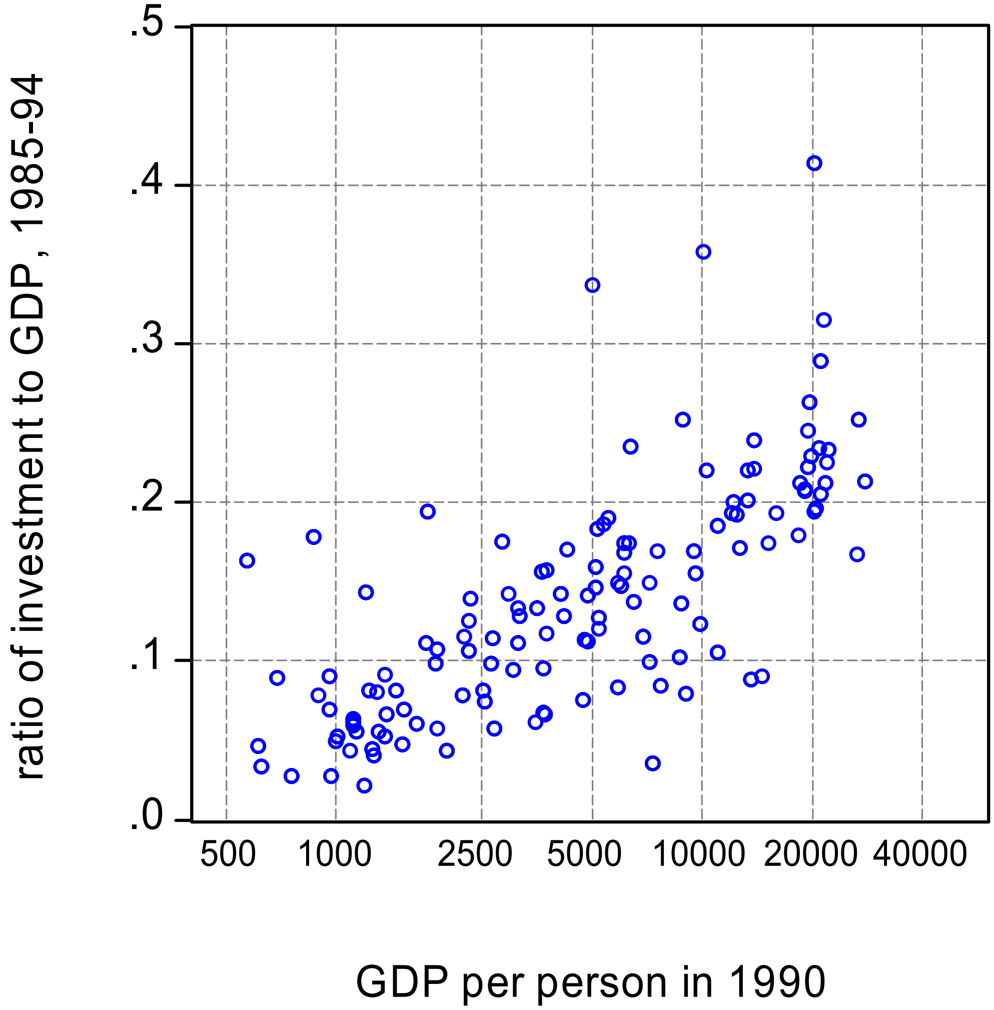


Figure 8.14

The Investment Ratio and GDP

For 135 countries, a higher per capita real GDP in 1990 (measured in 1995 U.S. dollars) matches up with a higher average ratio of investment to GDP from 1985 to 1994.

January 12, 2005

Chapter 9

Employment and Unemployment

The U.S. data show that labor input, L, is procyclical, that is, it moves during economic fluctuations in the same direction as real GDP. For example, in Figure 8.2, we measured L by total hours worked per year. This concept of L is strongly procyclical—from 1964.1 to 2004.3, the correlation of the cyclical part of total hours with the cyclical part of real GDP is 0.88. Moreover, this measure of L is nearly as volatile as real GDP.

In chapters 6 and 7, we analyzed economic fluctuations for given labor input, L. Thus, this analysis could not explain why L is procyclical. In chapter 8, we allowed L to change by introducing variable labor supply, L^s. Hence, this analysis might explain why L moves along with real GDP during economic fluctuations. To get this result, we relied on a positive response of the quantity of labor supplied, L^s, to the real wage rate, w/P. A high w/P in economic booms motivates households to supply more labor, and this response allows L to expand.

In the model, we assumed that the labor market always cleared, so that quantities supplied and demanded balanced:

$$L^s = L^d.$$

Total hours worked, L, then equaled the quantity supplied, L^s, and the quantity demanded, L^d.

Suppose, for the moment, that we neglect variations in hours worked per worker. In this case, each job would come with a standard number of hours of work per year. Changes in total hours worked would then reflect only changes in **employment**—the number of persons working. In this environment, we can think of the quantity of labor supplied, L^s, as the number of persons who wish to work, given the real wage rate, w/P, being offered. This measure of labor supply is called the **labor force.** We can think here of the quantity of labor demanded, L^d, as the number of jobs that employers want filled, given w/P.

In a market-clearing setting, the real wage rate, w/P, is determined, as usual, to equate the quantity of labor supplied, L^s, to the quantity demanded, L^d. Thus, the market-clearing employment, L, equals the labor force, L^s, and also equals the number of jobs that employers want filled, L^d. The real world departs from this environment in two major respects. First, the labor force is always greater than employment, and the difference between the two equals the number of persons **unemployed**. Second, the number of jobs that employers want filled is always greater than employment, and the difference between these two equals the number of job **vacancies**.

Economists usually focus on the labor-supply side of the labor market and therefore emphasize the labor force and the number of persons unemployed. One important statistic here is the **unemployment rate**, which equals the ratio of the number of persons unemployed to the labor force. Conversely, the **employment rate** is the ratio of the number of persons employed to the labor force. If we let u be the unemployment rate, we have

$$u = \textit{number unemployed / labor force}$$

$$= (\textit{labor force} - \textit{number employed}) \,/\, \textit{labor force}$$

$$= 1 - (\textit{number employed} \,/\, \textit{labor force})$$

$$= 1 - \textit{employment rate}.$$

Thus, if we rearrange terms, we have

$$\textit{employment rate} = 1\text{-}u.$$

We can also consider the demand side of the labor market. From this perspective, the **vacancy rate** is the ratio of the number of job vacancies to the total number of jobs that employers want filled. The employment rate from the standpoint of employers is the ratio of employment to the total number of jobs that employers want filled. One reason that this side of the market is less emphasized is that data on job vacancies are not as accurate as those on unemployment. Typically, economists have relied on incomplete information on **help-wanted advertising** in newspapers. Recently, however, the U.S. Bureau of Labor Statistics (BLS) has improved its measures of job vacancies.

From the labor-supply side, we can think of employment as determined by the following:

$$\textit{employment} = \textit{labor force} \cdot (\textit{employment} \,/\, \textit{labor force})$$

$$= \textit{labor force} \cdot \textit{employment rate}$$

$$= \textit{labor force} \cdot (1\text{-}u).$$

Our previous analysis assumed that the unemployment rate, u, was zero, so that variations in employment corresponded to variations in the labor force. Now, changes in employment can also reflect changes in u.

We saw in Figure 8.1 that employment is nearly as variable as real GDP—for 1954.1 to 2004.3, the standard deviations of the cyclical parts are 1.3% for employment and 1.6% for real GDP. In addition, these cyclical parts are strongly positively correlated—the correlation is 0.79—so that employment is clearly procyclical.

Figures 9.1 and 9.2 show how the two variables that determine employment—the labor force and the employment rate, $1\text{-}u$—contribute to the variations in U.S. employment from 1954.1 to 2004.3.[1] Figure 9.1 shows that the labor force is relatively stable—from 1954.1 to 2004.3, the standard deviation of the cyclical part is 0.4%. Moreover, the correlation with the cyclical part of real GDP is only 0.30. In contrast, the employment rate shown in Figure 9.2 is more variable—the standard deviation of the cyclical part is 0.7%—and is much more correlated with the cyclical part of real GDP—this correlation is 0.88. Thus, the procyclical variations in employment have a lot to do with changes in the employment rate and less to do with changes in the labor force. This finding means that our previous analysis—which focused on fluctuations in the labor force—is missing something important.

We can also add back the variations in hours worked per worker. We can write

total hours worked = employment ▪ (hours worked per worker).

We saw in Figure 8.2 that total hours worked is nearly as variable as real GDP—from 1964.1 to 2004.3, the standard deviation of the cyclical part is 1.5%, compared to 1.6% for real GDP. Total hours worked is also highly correlated with real GDP—the correlation between the cyclical parts is 0.88. Thus, total hours worked are more procyclical than employment.

Figures 9.1 and 9.2 considered the two variables that determine employment—the labor force and the employment rate. Figure 9.3 shows the additional variable, hours

[1] One problem is that the data on the labor force and the employment rate, shown in Figures 9.1 and 9.2, are from the survey of households, whereas the employment numbers in Figure 8.1 are from the survey of establishments (firms). This discrepancy does not affect our present discussion.

worked per worker, that determines total hours worked. From 1964.1 to 2004.3, the standard deviation of the cyclical part of hours worked per worker is 0.4% and the correlation with the cyclical part of real GDP is 0.73. Thus, from the standpoint of accounting for the overall cyclical fluctuations of total hours worked, hours worked per worker is less important than the employment rate and more important than the labor force.

Our previous market-clearing approach to the labor market is probably satisfactory for understanding the cyclical fluctuations in the labor force and in hours worked per worker. That is, in these cases, we can think of the real wage rate, w/P, as adjusting to equate the quantity of labor supplied, L^s, to the quantity demanded, L^d. However, this approach leaves unexplained the most important factor—the procyclical movements in the employment rate or, equivalently, the countercyclical movements in the unemployment rate.

To explain unemployment, we have to introduce some type of "friction" into the workings of the labor market. That is, we have to explain why persons in the labor force who lack jobs take some time to find them. Similarly, to understand job vacancies, we have to see why businesses with unfilled jobs take some time to fill them. Thus, the key to unemployment and vacancies is the process of persons searching for jobs and businesses searching for workers.

In our previous discussions of the labor market, we simplified by treating all workers and jobs as identical. However, in this world, the process of search among workers and businesses would be trivial. Thus, to make the analysis meaningful, we have to allow for differences among workers and jobs. Then we can think of the labor market as operating to find good matches between jobs and workers. Because jobs and workers differ, this matching process is difficult and time-consuming. Unemployment and vacancies arise as parts of this process.

The simple model of job matching developed in the next section has two main objectives. First, we want to explain why the levels of unemployment and vacancies are greater than zero. Second, we want to understand how unemployment and vacancies change over time. In particular, we want to explain why the employment rate is procyclical and the unemployment rate is countercyclical.

I. A Model of Job Finding

Consider a person who has just entered the labor force and is not yet employed. An example is a student who just graduated from school and is seeking his or her first job. Another example is someone who enters or reenters the labor force after raising a family. Suppose that this person searches for a position by visiting various firms. Each firm interviews job candidates to assess their likely qualifications for a position. As a result of each inspection, the firm estimates the candidate's potential marginal product, MPL. To keep things simple, assume that the firm offers the person a job with a real wage rate, *w/P*, equal to the estimated marginal product.[2] (We assume, only for simplicity, that each job entails a standard number of hours worked per week.)

[2] The marginal products of most people on most jobs would be negative. Consider, for example, how the typical college professor would perform as the chief executive officer of a major corporation, and vice versa. To capture this property in our model, we can imagine that firms make job offers with negative real wage rates. Since these offers would typically not be accepted, we can think equivalently of no job being offered in such cases.

The candidate decides whether to accept a job at the offered real wage rate, *w/P*. The alternative is to remain unemployed and continue to search for another job.[3] More search pays off if a subsequent wage offer exceeds the initial one. The cost of rejecting an offer is the wage income foregone while not working. This income must, however, be balanced against any income that people receive because they are unemployed. We denote by ω (the Greek letter omega) the effective real wage rate while unemployed. The amount ω includes **unemployment insurance** payments from the government and any value that persons attach to time spent not working (at "leisure"), rather than on the job.

In evaluating an offer, the first thing to consider is how it compares with others that are likely to be available. In making this comparison, a job seeker has in mind a probability distribution of real wage rates,[4] given the person's education, work experience, locational preferences, and so on. Figure 9.4 shows a possible distribution of offers. For each real wage rate, *w/P*, on the horizontal axis, the height of the curve shows the relative chance or probability of receiving that offer. For the curve shown, the offers usually fall in a middle range of *w/P*. There is, however, a small chance of getting either a very high offer (in the right tail of the distribution) or an offer near zero.

Figure 9.4 shows the value ω, which is the effective real wage rate received while unemployed. We know right away that a person would reject any job offer that paid less than ω. For the case shown in the figure, ω lies toward the left end of the distribution of real wage offers. This construction implies that most—but not all—offers exceed ω.

[3] We assume that it does not pay to accept a job and nevertheless keep searching. This assumption is reasonable in many cases because the costs of getting set up in a new job usually make it undesirable to take positions with short expected durations. Furthermore, it is likely to be easier to search for jobs while unemployed.

[4] We assume, for simplicity, that the attractiveness of a job depends only on the real wage rate. The basic results would not change if we took account of work location and working conditions, such as physical exertion, hours and timing of work, and pleasantness of surroundings.

Given the position of ω, the key decision is whether to accept a real wage offer, w/P, when it is greater than ω.

As mentioned before, a person may refuse a real wage rate, w/P, that exceeds ω to preserve the chance of getting a still better offer (see note 4). However, there is a trade-off because the job seeker then foregoes the real income $w/P - \omega$ while not working. The balancing of these forces generates what economists call a **reservation real wage** (or sometimes an acceptance real wage), denoted by $(w/P)'$. Offers below $(w/P)'$ are rejected, and those above $(w/P)'$ are accepted. If a job seeker sets a high value of $(w/P)'$, he or she will probably spend a long time (perhaps forever) unemployed and seeking work. In contrast, a low value of $(w/P)'$ (but still greater than ω) means that the expected time unemployed will be relatively brief. However, the expected real wage rate, w/P, received on the job will be lower the lower is $(w/P)'$.

The optimal reservation real wage, $(w/P)'$, depends on the shape of the wage-offer distribution in Figure 9.4, as well as the effective real wage while unemployed, ω, and the expected duration of a job.[5] For our purposes, we do not have to go through the details of the determination of the optimal $(w/P)'$. We can get the main results by describing the important properties that come out of this analysis.

First, because some job offers would generally be unacceptable—that is, $w/P < (w/P)'$ for some offers—it typically takes time for a job searcher to find an acceptable position. In the interim, the person is "unemployed," although actively engaged in job search. Thus, incomplete information about where to find the best job can explain positive amounts of unemployment.

---

[5] For a discussion of models of job search that involve a reservation wage, see Fleisher and Kniesner (1984, pp. 477-507).

Second, an increase in the effective real wage while unemployed, ω, motivates job seekers to raise their standards for job acceptance; that is, $(w/P)'$ increases. Therefore, for a given distribution of offers in Figure 9.4, it becomes more likely that $w/P < (w/P)'$ will apply. Hence, job offers will be rejected more often. It follows that job searchers tend to take longer to find a position when ω increases. For a group of persons, we therefore predict that a rise in ω reduces the **job-finding rate** and raises the expected **duration of unemployment**.

Third, suppose that the entire distribution of job offers becomes better. For example, a favorable technology shock might raise the marginal product, MPL, of all workers by 10%. Since each real wage offer, w/P, equals the value of a worker's potential marginal product, the distribution of real wage offers in Figure 9.4 shifts to the right. Specifically, the height of the new curve at each w/P equals the height of the original curve at $1.1\cdot(w/P)$. The main point is that the typical real wage offer, w/P, is now higher by 10%. Therefore, for a given reservation real wage, $(w/P)'$, job offers are more likely to lie in the acceptable range, where $w/P > (w/P)'$. Hence, the job-finding rate rises, and the expected duration of unemployment falls.

We have to consider, however, that a better distribution of real wage offers likely raises the reservation real wage, $(w/P)'$. We would expect $(w/P)'$ to increase if job seekers anticipate that the better distribution of real wage offers will apply in the future as well as the present. For example, a permanent improvement in technology would tend to raise potential marginal products, MPL, and, therefore, real wage offers permanently. In this case, $(w/P)'$ would rise. This increase in $(w/P)'$ works against our predicted rise in the job-finding rate. Specifically, in the example where all real wage offers, w/P, increase by 10%, the job-finding rate will rise only if the increase in *(w/P)′* is by less than 10%.

. There are two reasons why the increase in the reservation real wage, *(w/P)′*, will usually be smaller in proportion than the rise in the typical real wage offer, *w/P*. First, if the change in technology is less than permanent, future real wage offers will tend to rise by less than current offers. In this case, *(w/P)′* will also rise proportionately less than the real wage, *w/P*, offered currently to the typical job searcher.

Second, even if the improvement in technology and, hence, real wage offers is permanent, *(w/P)′* will rise by proportionately less than the typical real wage offer, *w/P*, if the effective real wage received while unemployed, ω, does not change. To see why, we can compare three scenarios, as follows.

- Scenario 1 is the initial situation, where offers of real wage rates, *w/P*, are given by the distribution in Figure 9.4 and the effective real wage received while unemployed is ω.
- Scenario 2 is the new situation, where the typical real wage offer, *w/P*, is permanently higher by 10%, and ω is unchanged.
- Scenario 3 is a hypothetical alternative where the typical real wage offer, *w/P*, is permanently higher by 10%, and ω is also permanently higher by 10%.

Compare scenario 1 with scenario 3. The only difference is that everything is scaled upward by 10% in scenario 3. Therefore, in weighing the trade-off between accepting or rejecting a job offer, it seems plausible (and is also correct) that a person would set the reservation real wage, *(w/P)′*, higher by 10% in scenario 3. But then the probability of accepting a job offer is the same in scenario 3 as in scenario 1. Hence, the job-finding rate is the same in these two cases.

Now compare scenario 3 with scenario 2. The only difference is that the real wage received while unemployed, ω, is higher by 10% in scenario 3. Therefore, a job seeker would set the reservation real wage, $(w/P)'$, lower in scenario 2 than in scenario 3. We conclude that the job-finding rate is higher in scenario 2 than in scenario 3. But we already found that scenario 3 has the same job-finding rate as scenario 1. Therefore, we have proven that the job-finding rate is higher in scenario 2 than in scenario 1. Thus, as claimed, a permanent improvement in real wage offers, w/P, raises the job-finding rate if ω is unchanged.

II. Search by Firms

Thus far, we have taken an unrealistic view of how firms participate in the job-search process. Firms received applications, evaluated candidates in terms of likely marginal products, MPL, and then expressed real wage offers, w/P, that equaled these marginal products. This model does not allow firms to utilize their information about the characteristics of jobs, the traits of workers who tend to be productive on these jobs, and the real wages that typically have to be paid for such workers. Firms would communicate this information by advertising job openings that specify ranges of requirements for education, work experience, and so on, and also indicate a salary range. Such advertisements appropriately screen out most potential applicants and tend to generate more rapid and better matches of workers to jobs.

Although search by firms is important in a well-functioning labor market, the allowance for this search does not change our major conclusions. In particular,

- It still takes time for workers to be matched with jobs, so that the expected durations of unemployment and vacancies are positive.
- An increase in workers' effective real wages while unemployed, ω, lowers the job-finding rate and raises the expected duration of unemployment.
- A favorable shock to productivity raises the job-finding rate and lowers the expected duration of unemployment.

III. Job Separations

Workers search for jobs that offer high real wages, and employers search for workers with high productivity. Although workers and firms evaluate each other as efficiently as possible, they often find out later that they made mistakes. An employer may learn, for example, that a worker is less productive than anticipated, or a worker may discover that he or she dislikes the job. When a job match looks significantly poorer than it did initially, firms are motivated to discharge the worker, or the worker is motivated to quit. These kinds of separations are more likely to occur the lower was the expected quality of the job match in the first place. That is, if the match was close to the margin of being mutually advantageous, relatively small changes in perceptions would be enough to trigger a separation.

Separations also arise because of changed circumstances, even when firms and workers accurately assessed each other at the outset. For example, an adverse shock to a firm's production function would lower a worker's marginal product, MPL, and lead, thereby, to a discharge. If we distinguish the goods produced by different firms, we would get a similar effect from a decline in the relative demand for a firm's goods. As in the previous paragraph, the tendency to separate is sensitive to how good the job match was at the outset. If the match was not expected to be that great, relatively small changes in production conditions or product demand would be sufficient to make the match mutually unattractive.

The cases just mentioned involve new information about the quality of a job match. However, separations also occur because jobs were known to be temporary at the outset. Examples include seasonal workers in agriculture or at sports stadiums.

On the other side of the labor market, workers experience changed circumstances with respect to family status, schooling, location, and retirement, as well as alternative job prospects. Some of these shifts are surprises, whereas others are predictable. The important point is that these changes induce workers to quit jobs.

We conclude that job separations take place for a variety of reasons. For a group of workers, we can identify factors that influence the **job-separation rate**. This rate is high, for example, among inexperienced workers who are hard to evaluate or for young persons who are likely to experience changes in family size or job preferences. The separation rate is also high in industries that are subject to frequent shocks to technology or product demand.

If there were no job separations, no new persons entering the labor force, and no new job positions, the search process would eventually eliminate unemployment and vacancies. But separations, new job seekers, and new job positions mean that the finding of jobs is continually offset by the creation of new unemployment and vacancies. We now illustrate this process with a simple example.

IV. Job Separations, Job Finding, and the Natural Unemployment Rate

In Figure 9.5, the box labeled L denotes the number of persons employed, and the box labeled U shows the number unemployed. To simplify, assume that the labor force, which equals $L + U$, does not change over time. (Remember that variations in the labor force are a relatively small part of the cyclical variations in employment.) Hence, we do not allow for retirements, entry of new persons into the labor force, or withdrawals of job seekers from the labor force. This example also does not consider the creation of new job positions by firms.

Because of the reevaluation of jobs and workers and because of changing circumstances for firms and workers, some fraction of those employed experience a job separation each period, say a month. In Figure 9.5, the arrow from L to U represents the number of job separations over a month. Since the labor force is constant, all those who lose jobs move from category L to category U. (We are being unrealistic here, because we are ignoring the possibility that a job loser finds a new job immediately and, therefore, never becomes unemployed.)

Since December 2000, the U.S. Bureau of Labor Statistics (BLS) has estimated the job-separation rate for the total non-farm economy. This rate is the ratio of total job separations over a month to total non-farm employment.[6] From December 2000 to August 2004, the separation rate averaged 3.2% per month. With a total non-farm payroll of 132 million persons (the number for September 2004), this separation rate

[6] The data are in the Job Openings and Labor Turnover Survey (or JOLTS), available from the BLS.

means that about 4.2 million job separations occur each month. This number is staggering—the U.S. job market has an enormous flow of persons out of jobs (and, as we shall see, also into jobs).

As discussed before, the other thing that happens is that unemployed persons find jobs. In Figure 9.5, the arrow pointing from U to L represents the number of unemployed persons who find jobs during a month. We can use BLS data on job-hiring rates to get a realistic number for the job-finding rate. From December 2000 to August 2004, the ratio of job hirings over a month to total non-farm employment averaged 3.3%—similar to the separation rate. In other words, while 4.2 million persons separated from their jobs each month, 4.3 million people (not necessarily the same persons) were hired each month. Thus, there are tremendous gross flows of persons out of and into employment. The difference between the two—job hirings less job separations—gives the net change in employment. For the U.S. economy from December 2000 to August 2004, this net change happened to be close to zero. The more general point is that the net change in employment is always very small in relation to the gross flows out of and into jobs.

To measure the job-finding rate, we have to express the number of jobs found in relation to the number of unemployed persons, rather than the number employed. If the unemployment rate, u, is 5.5% (the value in 2004), the number of unemployed persons corresponding to an employment level of 132 million is 7.7 million.[7] Therefore, the monthly job-finding rate implied by the BLS data on job turnover is $4.3/7.7 = 0.56$.[8] Thus, roughly half of those unemployed find a job within a month.

[7] We have $u = 0.055 = U/(132 + U)$, where 132 *million* $+ U$ is the labor force.. We can solve out to get $U = 7.26/0.945 = 7.7$ *million*.
[8] We are assuming here that all job-finding comes from the ranks of the unemployed, rather than from persons who come directly from other jobs or directly from outside of the labor force.

We can work through the process of job separation and job finding to determine the dynamics of persons employed and unemployed. To get a realistic example, we assume values for the job-separation and job-finding rates that correspond to the U.S. data. Specifically, we assume that the job-separation rate is 0.03 per month and the job-finding rate is 0.5 per month. To keep things simple, we assume that these rates are constant. This assumption is inaccurate because, during recessions, the job-separation rate tends to rise above average, and the job-finding rate tends to fall below average. The opposite pattern applies during booms.

Table 9.1 assumes that the labor force is fixed at 150 million people. (The actual number from the household survey toward the end of 2004 was 147 million.) Suppose that the economy starts in month 1 with an unemployment rate, *u*, of 10%. That is, we think of an economy that begins in a serious recession. Employment, *L*, starts at 135 million, and unemployment, *U*, starts at 15 million.

We can work through the process of job separation and finding to determine the time paths of employment and unemployment. In the first month, 3% of the 135.0 million employed—4.0 million persons—lose their jobs. At the same time, 50% of the 15.0 unemployed—7.5 million persons—find jobs. Hence, the net increase in employment during the month is 3.5 million—an enormous number that is not realistic.[9] Correspondingly, unemployment falls by 3.5 million. The unemployment rate declines accordingly to 7.7% (11.5/150).

As the number employed rises, job separations increase but only slightly—to 4.2 million in the second month. As the number unemployed falls, job findings fall—to 5.8

[9] The results are unrealistic because we assumed constant rates of job separation and job finding. In a recession—which is the starting point for our example—the job-separation rate would be higher than average, and the job-finding rate would be lower than average.

million in the second month. Therefore, in month 2, employment rises on net by 1.6 million, and unemployment falls by 1.6 million. The unemployment rate is now down to 6.6% (9.9/150).

This process continues until the numbers of job separations and findings are the same. In our example, the economy gets close to this balance in month 6—when employment reaches 141.5 million and unemployment equals 8.5 million. The corresponding unemployment rate is 5.7% (8.5/150). Therefore, in this model, we can say that the **natural unemployment rate** is 5.7%. The economy tends toward this rate automatically, given the assumed (constant) rates at which people lose and find jobs.

This model, although not fully realistic, brings out some important points about the natural unemployment rate. First, although the unemployment rate eventually stays constant at the natural rate, there is still a large amount of job turnover. In the model (and also in the U.S. data), about 4 million people lose and find jobs each month when the unemployment rate equals its natural value of 5.7%. These large flows between employment and unemployment are a normal part of the operation of a fluid labor market.

Second, the keys to the dynamics of employment and unemployment are the rates of job separation and finding. In our example, we assumed that these rates were fixed at 3% and 50%, respectively. Our earlier discussion suggested that these rates would depend on characteristics of workers and jobs. For example, we discussed effects from a person's age and job experience, from the effective real wage paid while unemployed, ω, and from the variability of an industry's supply and demand conditions. The rates of job separation and finding depend also on shifts to economy-wide productivity, for example, the disturbances that we previously called technology shocks.

The rates of job separation and finding determine the natural unemployment rate—5.7% in the example of Table 9.1. That value corresponds to a monthly job-separation rate of 3% and a monthly job-finding rate of 50%. To see how the natural unemployment rate is determined more generally, let σ (the Greek letter sigma) be the job-separation rate and φ (the Greek letter phi) the job-finding rate. The change in the number of persons employed over a month, ΔL, is given by

$$(9.1) \qquad \Delta L = \varphi U - \sigma L$$
$$= \text{job findings} - \text{job separations}.$$

Note that the first term, φU, is the number of unemployed persons who find jobs over a month, and the second term, σL, is the number of employed persons who lose jobs over a month.[10]

Equation (9.1) implies that employment, L, increases and unemployment, U, decreases if job findings, φU, are greater than job separations, σL. In the reverse case, L decreases and U increases. To determine the long-run levels of L and U, we have to find the situation where L and U are constant. This constancy requires $\Delta L = 0$, and equation (9.1) says that $\Delta L = 0$ when job findings equal job separations:

$$\varphi U = \sigma L.$$

To solve for the long-run levels of L and U, we have to use our assumption that the labor force, $L + U$, is constant at 150 (million). In this case, we can substitute $L = 150 - U$ in the last equation to get

$$\varphi U = \sigma \cdot (150 - U).$$

If we combine the two terms involving U and put them on the left-hand side, we get

[10] The approximation here is that a person who finds a job is assumed to stay employed for at least a month and that a person who becomes unemployed is assumed to take at least a month to find a job.

$$U \cdot (\varphi + \sigma) = 150 \cdot \sigma.$$

Therefore, the long-run number unemployed is given by

$$U = 150 \cdot \sigma/(\varphi + \sigma).$$

We can then compute the natural unemployment rate, $u = U/150$, as

Key equation (natural unemployment rate):

(9.2) $$u = \sigma/(\varphi + \sigma).$$

Equation (9.2) relates the natural unemployment rate, u, to the rates of job separation, σ, and job finding, φ.[11] A higher rate of job separation, σ, raises u, and a higher rate of job finding, φ, lowers u. Thus, to think about the determinants of the unemployment rate, we have to consider factors that influence σ and φ. For example, since an increase in the effective real wage paid while unemployed, ω, lowers φ, we predict an increase in u. Thus, more generous unemployment insurance programs would raise the long-run average unemployment rate. We might also think that the Internet assists in the process of job matching and, thereby, raises φ. Hence, our prediction is that the invention of the Internet would lower the natural unemployment rate, u.

V. Employment and Unemployment during a Recession

We showed in chapters 7 and 8 that shifts to the production function—technology shocks—could account for some characteristics of real-world economic fluctuations. In particular, an adverse shock could generate a recession that featured declines in real GDP and labor input. Now we can use the extensions in this chapter to show how technology

[11] For more thorough analyses of this type of model, see Hall (1979), Ramaswami (1983), and Darby, Haltiwanger, and Plant (1985).

shocks affect employment and unemployment. We consider the case of a fixed labor force, so that the change in unemployment is the negative of the change in employment.

Suppose that the economy begins in the long-run situation described in Table 9.1. In this setting, the job-separation rate, σ, equals 0.03 per month and the job-finding rate, φ, equals 0.50 per month. If the labor force is constant at 150 million, we have that employment, L, equals 141.5 million and unemployment, U, equals 8.5 million. Although L and U are constant, the labor market features job losses and job findings of 4.2 million per month.

Suppose that an adverse shock reduces the marginal product of labor for the typical worker and job. One effect, discussed earlier, is that the job-finding rate, φ, declines. The reason is that market opportunities have become poorer—probably temporarily—relative to the effective real wage received while unemployed, ω. For the same reason, existing job matches become mutually less attractive for firms and workers. Therefore, job separations tend to increase, especially in the forms of layoffs and firings. Hence, the job-separation rate, σ, increases. For the purposes of an example, suppose that φ falls from 0.50 to 0.40, and σ rises from 0.03 to 0.04.

To see the dynamic effects on employment, L, and unemployment, U, suppose that the labor force is fixed at 150 million and that L and U start at their long-run values from Table 9.1 of 141.5 million and 8.5 million, respectively. Table 9.2 shows that L falls gradually and U rises gradually over time. For example, in month 1, although 3.4 million people still find jobs, they are outnumbered by the 5.7 million who lose jobs. Hence, employment falls during the month by 2.3 million, and unemployment expands correspondingly.

The assumption in Table 9.2 is that the job-separation and job-finding rates, σ and φ, return to their normal values in month 4. In response, L and U return gradually to their long-run values. Since σ and φ now take on the same values as those assumed in Table 9.1, the economy returns eventually to the same long-run position, where L = 141.5 million and U = 8.5 million.

Although the numerical example in Table 9.2 is too simple to be fully realistic, it does bring out a number of features of real-world recessions. First, the buildup of a recession involves a period of gradually falling employment and gradually rising unemployment. Second, even after an economic recovery begins, it takes a substantial period for employment and unemployment to return to their pre-recession levels. Third, even during a recession, substantial numbers of jobs are created—they are just outnumbered by the jobs lost.

VI. Vacancies

We can extend the model to allow for job vacancies, which are posted jobs that have not yet been filled. Suppose that firms have some idea about potential workers in the sense of their marginal products, MPL. Firms also have a sense of the real wage rate, w/P, that has to be offered to induce the typical worker to accept a job. Finally, assume that there are costs for posting job openings and for interviewing potential workers. Given these considerations, firms would determine how many job openings to advertise.

For our purposes, we do not have to work through the details of a model that determines the optimal number of job openings. We can just note some important properties that would emerge from such a model. One result is that an increase in prospective marginal products, MPL, would raise the number of job openings. A second result is that an increase in the real wage rate, *w/P*, required to get workers to accept jobs would lower the number of job openings. Finally, a reduction in costs of posting jobs and processing applications—occasioned, for example, by the rise of the Internet—would raise the number of job openings.

In previous discussions, we stressed the effects from technology shocks. For example, a favorable shock raises the marginal product of labor, MPL. For given real wage rates, *w/P*, this type of shock would motivate firms to post more job openings. Therefore, we predict that job vacancies would increase. Conversely, an unfavorable shock would reduce job vacancies. Thus, our prediction is that job vacancies would be procyclical—high in booms and low in recessions.

As already mentioned, data on U.S. job vacancies have been inadequate until the introduction of the JOLTS survey by the Bureau of Labor Statistics in December 2000. To fill this gap in the data, economists have used as a proxy for job vacancies an index of **help-wanted advertising** in major newspapers.[12] However, this index has a number of shortcomings. First, it does not cover all job vacancies and is not representative of job vacancies in the entire country. Second, the role of newspaper advertising in the job-search process has shifted over time. In particular, the growth of the Internet in recent years has diminished the importance of newspaper advertising. Despite these problems, we can probably use the cyclical part of the help-wanted advertising index to get a sense of the short-term fluctuations of job vacancies.

Figure 9.6 shows the cyclical pattern for the log of the help-wanted index from 1954.1 to 2004.3. The index is highly volatile—the standard deviation of the cyclical

[12] This index is compiled by the Conference Board. For a discussion, see Abraham (1987).

part is 0.14 compared to only 0.016 for real GDP. We do find, as predicted, that the index is procyclical—the correlation of the cyclical part with the cyclical part of real GDP is 0.91. Thus, these data confirm the prediction that job vacancies are procyclical.

The procyclical pattern for vacancies helps to explain why the job-finding rate would be high in a boom and low in a recession. In our previous analysis, we argued that job seekers would be more likely to accept job offers in a boom because of the increase in the real wage rate, *w/P*. The increase in vacancies reinforces this response because the greater availability of jobs makes it easier for workers to find positions that look like good matches. Thus, in a boom, the rate of job acceptance increases partly because the pay is better and partly because attractive positions are easier to find. Conversely, in a boom, the rate of job acceptance decreases because the pay is worse and because attractive positions are harder to find.

We know from Figure 9.2 that the employment rate is procyclical and, hence, that the unemployment rate (which equals one minus the employment rate) is countercyclical. Thus, in a boom, vacancies tend to rise, and unemployment tends to fall. The opposite pattern holds for recessions. In other words, during economic fluctuations, vacancies and unemployment move in opposite directions. Economists often express this relation in the form of a scatter plot between the cyclical parts of the unemployment rate and a measure of job vacancies.[13] Figure 9.7 provides this scatter plot for U.S. data from 1954.1 to 2004.3. The horizontal axis has the cyclical part of the unemployment rate, *u*. The vertical axis has the cyclical part of the log of the help-wanted advertising index. Note the clear downward slope—the correlation between the two series is -0.93. Thus, the

[13] This type of plot is called a Beveridge curve, named after the British economist William Beveridge. For further discussion, see Shimer (2003).

data strongly confirm that a low unemployment rate matches up with high vacancies, whereas a high unemployment rate matches up with low vacancies.

Seasonal Fluctuations

Our analysis views economic fluctuations as resulting from shocks to the technology level, *A*. As mentioned before, these kinds of shocks are the focus of real business cycle theory. We also observed that shocks other than literal shifts to technology, such as harvest failures and strikes, have effects that resemble those from changes in *A*. Seasonal fluctuations turn out to be additional sources of variation that resemble shifts to the technology.

As noted in chapter 2, economists usually use seasonally-adjusted data (provided by the Bureau of Economic Analysis) to study economic fluctuations. The seasonal-adjustment procedure eliminates the normal fluctuation of a variable, such as real GDP, from winter (the first quarter) to spring (the second quarter), and so on. In particular, in the raw data, real GDP tends to rise each year toward a peak in the fourth quarter. This systematic quarter-to-quarter pattern does not appear in the seasonally-adjusted data.

Although most economists focus on seasonally-adjusted data, Robert Barsky and Jeff Miron (1988) reached some interesting conclusions by studying the raw, unadjusted numbers. First, the seasonal fluctuations in quantities—real

GDP, consumption, investment, employment, and unemployment—are larger than the variations associated with typical recessions and booms. From 1948 to 1985, over 80% of the quarterly fluctuations in real GDP and over 60% of those in the unemployment rate were due to systematic seasonal factors (Barsky and Miron [1988, Table 1]). Furthermore, the seasonal patterns of co-movement among real GDP and its major components and between real GDP and employment look similar to the patterns found in economic fluctuations (*ibid*., Table 2). In the seasonal pattern, for example, investment and consumption tend to move along with real GDP, and investment is far more volatile than consumption. Jeff Miron (1988) shows that these findings for the United States apply also to a sample of 25 industrialized or semi-industrialized countries.

Seasonal fluctuations presumably reflect influences of weather and holidays. We can think of some of these factors as regular variations in technology, such as the effect of weather on the construction industry. Other factors correspond to variations in household preferences, such as the positive impact of Christmas on consumer demand and the negative effect of summer vacations on labor supply. Thus, the seasonal factors resemble the disturbances stressed in real business cycle theories—shocks to technology, which we have emphasized, and to preferences, which are considered in some real business cycle models. The magnitude of the seasonal fluctuations shows that these kinds of disturbances can be quantitatively important in the short run. That is, the seasonal evidence weakens the argument made by some economists that shocks to technology and preferences are not large enough to account for the observed magnitude of recessions and booms.

VII. Summing Up

In the U.S. data, labor input—measured, for example, by total worker-hours—is strongly procyclical. Worker-hours can be broken down into three components: the labor force, the employment rate (the fraction of the labor force with jobs), and the average hours worked per worker. The most important part of the cyclical fluctuations in worker-hours turns out to be the employment rate. Changes in average hours worked come second in importance, and movements in the labor force come third. Our analysis in chapter 8 applies to the last two of these components but does not explain the most important part, the variations in the employment rate.

This chapter improves the model by adding an analysis of job finding and job separation. The new analysis explains, first, why the unemployment rate would be greater than zero, that is, why the employment rate would be less than 100%. Second, the analysis shows why the unemployment rate would be countercyclical, so that the employment rate would be procyclical. Hence, the extended model provides a better explanation of the procyclical behavior of labor input.

Questions and Problems

Mainly for review

9.1. What is the definition of the unemployment rate? Since it does not include persons who moved from "unemployed" to "out of the labor force," does it underestimate the true unemployment rate? Can you think of reasons why the reported numbers may overestimate the true unemployment rate?

9.2. Suppose that a job seeker receives a real wage offer, w/P, that exceeds his or her effective real wage rate, ω, received while unemployed. Why might the person reject the offer?

9.3. Once a job seeker and a firm find a job match, why might they choose subsequently to end the match? List some influences on the rate of job separation.

9.4. What is the natural rate of unemployment? Why might the unemployment rate differ from the natural rate? Can the natural rate of unemployment change over time?

Problems for discussion

9.x. The job-finding rate

Discuss the effects on the job-finding rate and the expected duration of unemployment from the following:

a. an increase in the amount of unemployment-insurance benefits.

b. an increase in the allowable duration of unemployment-insurance benefits.

c. a technological change, such as the Internet, that improves the matching of workers and jobs.

9.x. The job-finding rate, the job-separation rate, and the dynamics of the unemployment rate

Suppose that the labor force has 100 million people, of whom 92 million initially have jobs and 8 million are unemployed. Assume that the job-separation rate is 1% per month and the job-finding rate is 20% per month. Also, assume that we can neglect movements in and out of the labor force. Trace out the time path of employment and unemployment. What is the natural unemployment rate?

9.x. Cyclical behavior of the labor force

Figure 9.1 shows that the labor force is weakly procyclical. What pattern would you predict on theoretical grounds? (Hint: Think first about people's incentives to leave the labor force—that is, stop looking for work—during a recession. Are there also incentives for people to enter the labor force during recessions?)

9.x. Job vacancies

Suppose that economic fluctuations are caused by shocks to the technology level, A. What do you predict for the cyclical behavior of job vacancies? How then would fluctuations in vacancies relate to fluctuations in the unemployment rate? How does your answer relate to the Beveridge curve shown in Figure 9.7?

Table 9.1 Dynamics of Unemployment and the Natural Rate of Unemployment							
	start of month			**during month**			
Month	**Number Employed (*L*) (millions)**	**Number Unemployed (*U*) (millions)**	**Unemploy-ment rate (*u*)**	**Number who lose jobs (millions)**	**Number who find jobs (millions)**	**Change in *L* (ΔL) (millions)**	**Change in *U* (ΔU) (millions)**
1	135.0	15.0	0.100	4.0	7.5	3.5	-3.5
2	138.5	11.5	0.077	4.2	5.8	1.6	-1.6
3	140.1	9.9	0.066	4.2	5.0	0.8	-0.8
4	140.9	9.1	0.061	4.2	4.6	0.4	-0.4
5	141.3	8.7	0.058	4.2	4.4	0.2	-0.2
6	141.5	8.5	0.057	4.2	4.2	0.0	0.0
∞	141.5	8.5	0.057	4.2	4.2	0.0	0.0

Note: We assume that the economy starts with 135 million people employed, *L*, and 15 million unemployed, *U*. The labor force, $L + U$, is assumed to be constant at 150 million. The unemployment rate is $u = U/(L+U) = U/150$. From Figure 9.5, 3% of those employed lose jobs each month, and 50% of those unemployed find jobs. The net change in employment, ΔL, is therefore $0.5 \cdot U - 0.03 \cdot L$. The net change in unemployment, ΔU, is the negative of ΔL. When *L* reaches 141.5 million and *U* reaches 8.5 million, ΔL and ΔU equal zero. Thus, the natural unemployment rate is $u = 8.5/150 = 5.7\%$.

Table 9.2 Dynamics of Employment and Unemployment during a Recession							
	start of month		**during month**				
Month	**Number employed (*L*) (millions)**	**Number unemployed (*U*) (millions)**	**Job-separation rate (σ) (monthly)**	**Job-finding rate (φ) (monthly)**	**Number who lose jobs (millions)**	**Number who find jobs (millions)**	**Change in *L* (ΔL) (millions)**
1	141.5	8.5	0.04	0.40	5.7	3.4	-2.3
2	139.2	10.8	0.04	0.40	5.6	4.3	-1.3
3	137.9	12.1	0.04	0.40	5.5	4.8	-0.7
4	137.2	12.8	0.03	0.50	4.1	6.4	2.3
5	139.5	10.5	0.03	0.50	4.2	5.2	1.0
6	140.5	9.5	0.03	0.50	4.2	4.8	0.6
7	141.1	8.9	0.03	0.50	4.2	4.4	0.2
8	141.3	8.7	0.03	0.50	4.2	4.4	0.2
9	141.5	8.5	0.03	0.50	4.2	4.2	0.0
∞	141.5	8.5	0.03	0.50	4.2	4.2	0.0

Note: The economy starts from the long-run position of Table 9.1, with employment, *L*, of 141.5 million and unemployment, *U*, of 8.5 million. These values correspond to a job-separation rate, σ, of 0.03 per month and a job-finding rate, φ, of 0.50 per month. In month 1, the start of a recession raises σ to 0.04 and lowers φ to 0.40. These values generate a gradual decline in *L* and a gradual rise in *U*. In month 4, we assume that σ returns to 0.03 and φ to 0.50. These values generate a gradual return of *L* toward 141.5 million and of *U* toward 8.5 million.

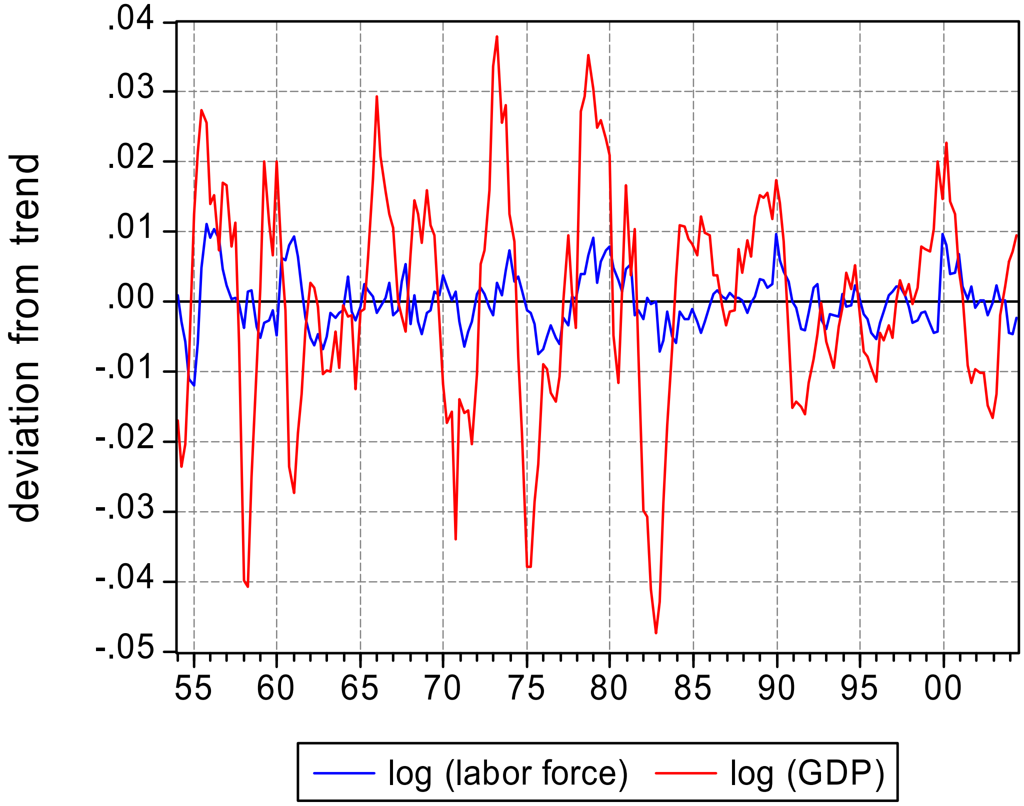


Figure 9.1

Cyclical Parts of U.S. GDP and Labor Force

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of the labor force from its trend. The labor force—the number of persons employed or seeking employment—comes from the BLS Household Survey. The data on real GDP and the labor force are quarterly and seasonally adjusted. (The underlying data on the labor force are monthly.) The labor force is weakly procyclical—it fluctuates weakly with real GDP and is less variable than real GDP.

Figure 9.2

Cyclical Parts of U.S. GDP and the Employment Rate

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the employment rate from its trend. The employment rate is the ratio of the number employed to the labor force. (The measures of employment and labor force come from the BLS Household Survey.) The data on real GDP and the employment rate are quarterly and seasonally adjusted. (The underlying data on the employment rate are monthly.) The labor force is strongly procyclical—it fluctuates in the same direction as real GDP.

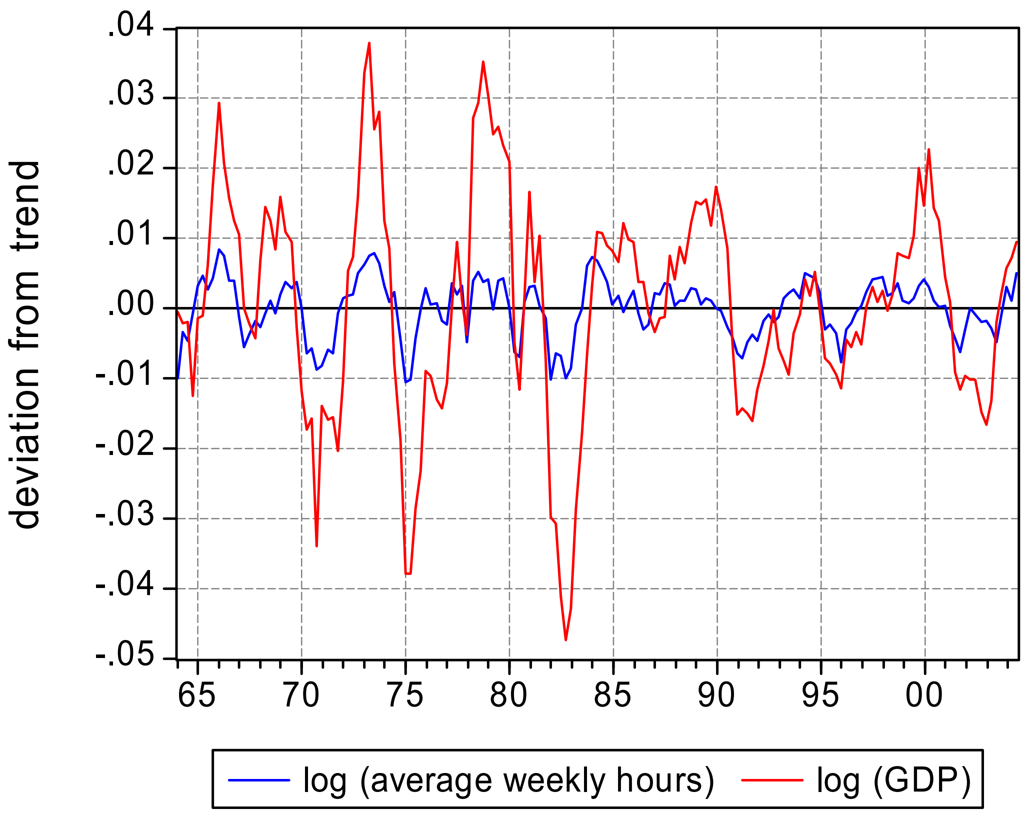


Figure 9.3

Cyclical Parts of U.S. GDP and Average Weekly Hours

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of average weekly hours worked from its trend. The data for weekly hours come from the BLS payroll survey and refer to the private, non-agricultural economy. These numbers are available since 1964. The data on real GDP and weekly hours are quarterly and seasonally adjusted. (The underlying data on weekly hours are monthly.) Average weekly hours are procyclical—they fluctuate with real GDP but are less variable than real GDP.

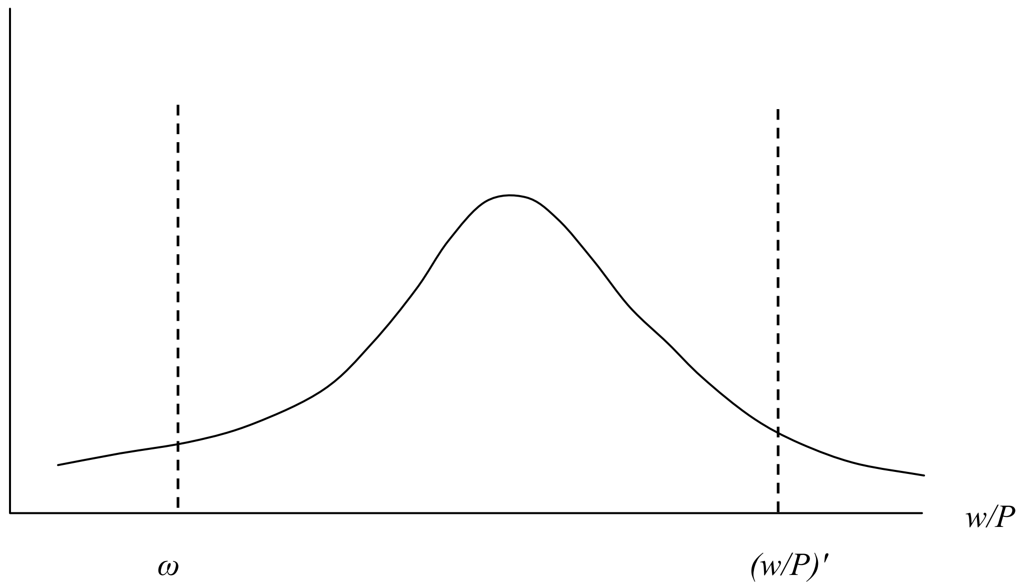


Figure 9.4

Distribution of Real Wage Offers

The curve shows the chances of receiving real wage offers, *w/P*, of different sizes. The higher the curve, the more likely that real wage offers of that size will be received. Note that ω is the effective real wage received while unemployed, and *(w/P)′* is the reservation real wage. Job offers are accepted if they are at least as good as *(w/P)′* and are otherwise rejected.

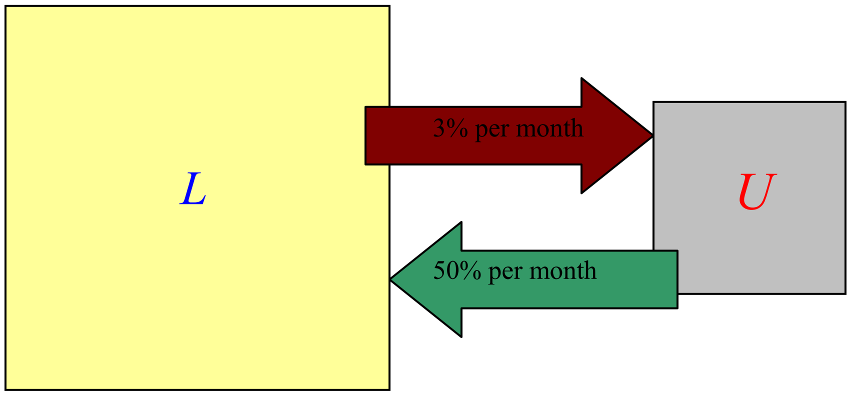


Figure 9.5

Movements between Employment and Unemployment

In this example, 3% of those employed, L, lose their jobs each month and 50% of those unemployed, U, find jobs. The net change in employment, ΔL, is therefore $0.5 \cdot U - 0.03 \cdot L$. The change in unemployment, ΔU, is the negative of ΔL.

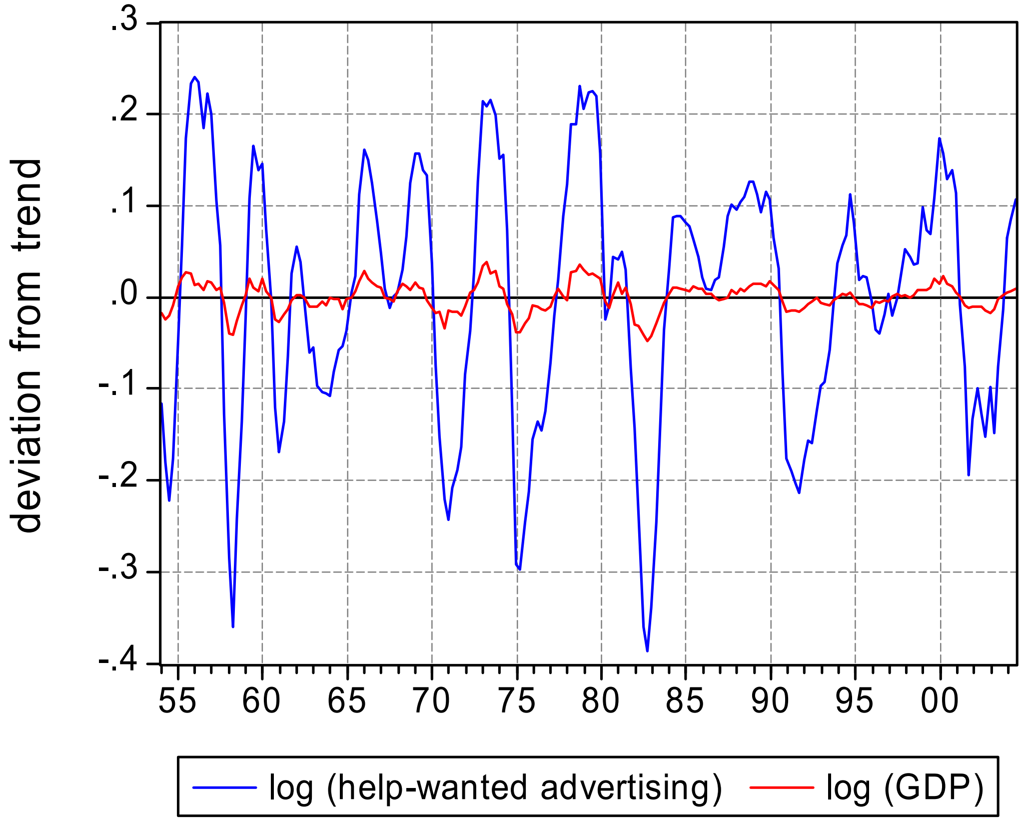


Figure 9.6

Cyclical Parts of U.S. GDP and Help-Wanted Advertising

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of the help-wanted-advertising index from its trend. The data on help-wanted advertising in major newspapers come from the Conference Board. The data on real GDP and help-wanted advertising are quarterly and seasonally adjusted. (The underlying data on help-wanted advertising are monthly.) The help-wanted-advertising index is procyclical—it fluctuates with real GDP and is more variable than real GDP.

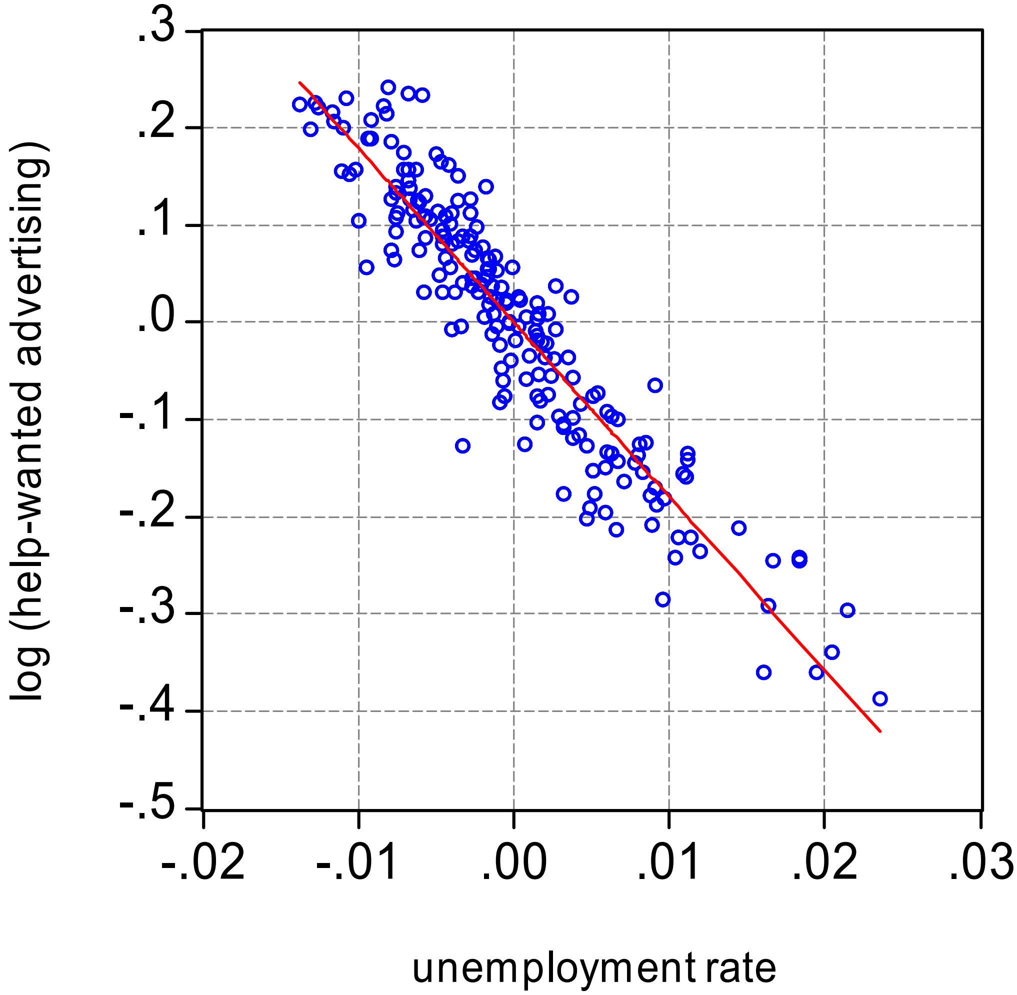


Figure 9.7

The Unemployment Rate and Help-Wanted Advertising (a Beveridge Curve)

The horizontal axis plots the cyclical part of the unemployment rate. This variable is the same, except for a minus sign, as the cyclical part of the employment rate, shown in Figure 9.2. The vertical axis has the cyclical part of the log of the index of help-wanted advertising (from Figure 9.6). The plots use quarterly, seasonally-adjusted data from 1954.1 to 2004.3. (The underlying data on both variables are monthly.)

January 12, 2005

Chapter 10

The Demand for Money and the Price Level

Our model has three forms of assets: money, bonds, and ownership of capital. So far, we have not analyzed how much money households hold or how these holdings change over time. We therefore carried out our analysis in chapters 6-8 under the assumption that each household held a constant stock of money, M. Now we extend the microeconomic foundations of the model to explain why households hold part of their assets as money, that is, we explain the **demand for money**.

We have already assumed that money is the sole **medium of exchange** in the economy. Households exchange money for goods on the goods market, money for labor on the labor market, money for capital services on the rental market, and money for bonds on the bond market. However, households do not directly exchange goods for goods (a process called **barter**) or bonds for goods, and so on.

The money in our model matches up with **paper currency** issued by a government. For example, the money could be U.S. dollar notes issued by the Federal Reserve, euro notes issued by the European Central Bank, and almost 200 other forms of paper currency issued by the world's governments. These paper currencies are sometimes called **fiat money** because they have value due to government fiat, rather than through intrinsic value. At earlier times, societies tended to rely more on **commodity money**, such as gold and silver coins, which do have intrinsic value. The coins are valued, at least in part, for their content of gold or silver. In the nearby box, we discuss how another commodity, the cigarette, served as money in a prisoner-of-war camp. In our model, money has no intrinsic value; it is just a piece of paper issued by the government. Therefore, we do not have to consider any resources used up when intrinsically valuable goods serve as money.

Money in a prisoner-of-war camp

R.A. Radford (1945) described his experience with the economy of a German prisoner-of-war camp during World War II. He observed that cigarettes became the primary medium of exchange, with many goods exchanging for cigarettes, which were then used to buy other goods. In addition, most prices were expressed in terms of cigarettes, for example, as four cigarettes per ration of treacle (a form of syrup).

Radford noted that cigarettes had several attractive characteristics as money: "homogeneous, reasonably durable, and of convenient size for the smallest or, in packets, for the largest transactions." One drawback, applicable also to other commodity moneys, was the resource cost from using cigarettes as a medium of exchange. That is, the cigarettes used as money could not simultaneously be smoked and, worse yet, might deteriorate physically over time.

Radford discussed an attempt to introduce a paper money as an alternative medium of exchange. This money was issued by the camp restaurant and was supposed to be redeemable for a fixed quantity of food. However, problems arose with respect to the credibility of the promised food value of the paper money, and cigarettes remained as the primary medium of exchange. For our purposes, the most interesting lesson from Radford's story is that a medium of exchange is important in any economy, even a P.O.W. camp.

If we think of money as paper currency issued by the government, there are several reasons why this money might occupy the dominant position as an economy's medium of exchange. First, the government may impose legal restrictions that prevent private parties, such as Microsoft Corporation, from issuing small-size, interest-bearing bonds that could serve conveniently as hand-to-hand currency. Further, the government may enact statutes that reinforce the use of its money. As an example, there is the proclamation that the U.S. dollar is "**legal tender** for all debts public and private."[1] Also, U.S. courts are more inclined to enforce contracts that are denominated in U.S. dollars rather than in some other unit.

Another consideration is the cost of establishing one's money as reliable and convenient. These costs include prevention of counterfeiting, replacement of worn-out notes, willingness to convert notes into different denominations, and so on. Because of these costs, money would always tend to bear interest at a rate lower than bonds. In fact,

[1] Note that this provision does not specify the price, P, at which currency exchanges for goods. If P were infinite, what would the legal-tender property mean?

because of the inconvenience of paying interest on hand-to-hand currency, the interest rate on currency is typically zero. That is, if one holds \$1 of currency and does not lose it, one still has \$1 of currency in the future.

We can relate our abstract concept of money to conventional measures of the money stock. The theoretical construct corresponds most closely to currency held by the public. (This definition excludes currency held in vaults of banks and other depository institutions.) In September 2004, the amount of (seasonally-adjusted) currency held by the public in the United States was \$692 billion, which amounted to 5.9% of nominal GDP. This amount of currency is surprisingly large—about \$2400 for each person in the United States. In the nearby box, we note that much of the currency is in \$100 bills, many of which are held abroad.

Where is all the currency?

In September 2004, the amount of U.S. currency held by the public was \$692 billion or about \$2400 per person in the United States. To understand this surprisingly large number, one can start with the observation that around 70% of the currency by value is in \$100 bills. (The data on currency by denomination are in the *Treasury Bulletin*.) Thus, much of the currency is likely not used for ordinary transactions. Because currency is anonymous, it is attractive for illegal activities, such as the drug trade. Currency transactions also facilitate tax evasion. However, the amounts of U.S. currency held for these purposes are unknown.

More is known about the amounts of U.S. currency held abroad, mostly in the form of $100 bills. Foreigners like U.S. money as a store of value and a medium of exchange because the money has a reasonably stable value and can readily be exchanged for goods or other assets. In addition, transactions carried out in currency can usually be hidden from local governments, and this secrecy is especially attractive when the government is oppressive. The foreign demand for U.S. currency is especially high in countries experiencing economic and political turmoil. A recent joint study by the Federal Reserve and the U.S. Treasury estimated that 55-60% of the total of U.S. currency in 2002 was held abroad. The geographical division was estimated to be 25% in Latin America (with Argentina the highest demander), 20% in the Middle East and Africa, 15% in Asia, and 40% in Europe (with Russia and other former Soviet republics as particularly high users). For additional discussion, see Richard Porter and Ruth Judson (1996) and Board of Governors of the Federal Reserve System (2003).

The term "money" typically refers to a **monetary aggregate** that is broader than currency. The most common definition, called **M1**, attempts to classify as money the assets that serve regularly as media of exchange. This concept adds to currency held by the public the **checkable deposits** issued by banks and other financial institutions. The amount of these checkable deposits (including a relatively small amount of travelers' checks) in the United States in September 2004 was $653 billion, or 5.6% of nominal

GDP.[2] Therefore, M1—the sum of currency and checkable deposits—was $1345 billion in September 2004, or 11.5% of nominal GDP. This total of U.S. M1 was 51% in currency and 49% in checkable deposits. At earlier times, relatively less of M1 was in currency and relatively more was in checkable deposits. For example, in 1960, only 19% of M1 was in currency, whereas 81% was in checkable deposits.

Table 10.1 shows ratios of currency to nominal GDP for OECD countries (plus China) in 1960, 1980, and 2000. Notice that the ratio of currency to GDP declined over time in most countries—a typical case, for France, showed a decrease from 0.133 in 1960 to 0.052 in 1980 and 0.035 in 2000. However, in some countries, the ratio leveled off or even rose from 1980 to 2000—this pattern applied to Canada, Finland, Germany, Japan, Spain, and the United States. In 2000, the highest currency ratio was 0.121 in Japan, and the lowest was 0.019 in New Zealand. The United States, at 0.059, was close to the median. Table 10.2 shows comparable figures with money defined to be M1.

Still broader definitions of money add in other kinds of deposits held at financial institutions. For example, **M2** ($6330 billion in the United States in September 2004) includes household holdings of savings deposits, time deposits, and retail money-market mutual funds. An even broader aggregate, **M3** ($9312 billion in September 2004), adds in institutional money-market funds, large time deposits, repurchase agreements, and Eurodollar accounts. However, the M2 and M3 definitions go beyond the concept of money as a medium of exchange. In our model, it is best to identify money with currency held by the public.

---

[2] The standard definition of checkable deposits includes travelers' checks issued by banks. Non-bank travelers' checks, amounting to $7.6 billion in September 2004, are included in M1 but not in the usual measure of checkable deposits. The figure of $653 billion for checkable deposits includes the $7.6 billion of non-bank travelers' checks.

Table 10.1			
Ratios of Currency to Nominal GDP			
Country	**1960**	**1980**	**2000**
Australia	0.054	0.036	0.041
Austria	0.119	0.078	0.071
Belgium	0.220	0.110	0.054
Canada	0.046	0.034	0.034
China	--	--	0.072
Denmark	0.068	0.032	0.029
Finland	0.036	0.025	0.025
France	0.133	0.052	0.035
Germany	0.072	0.062	0.070
Greece	0.103	0.130	--
Ireland	0.117	0.077	0.052
Italy	--	0.070	0.066
Japan	0.069	0.072	0.121
Netherlands	0.125	0.064	0.047
New Zealand	0.061	0.025	0.019
Norway	0.112	0.060	0.030
Portugal	0.177	0.131	0.057
South Korea	0.059	0.049	0.034
Spain	0.120	0.083	0.099
Sweden	0.090	0.064	0.043
Switzerland	0.197	0.141	0.093
United Kingdom	0.081	0.044	0.025
United States	0.056	0.042	0.059

Note: The table shows the ratio of currency held by the public to nominal GDP. The data are from International Monetary Fund, *International Financial Statistics*.

Table 10.2			
Ratios of M1 to Nominal GDP			
Country	**1960**	**1980**	**2000**
Australia	0.228	0.126	0.211
Austria	0.197	0.151	0.280
Belgium	0.322	0.192	0.271
Canada	0.152	0.112	0.213
China	--	--	0.146
Denmark	0.246	0.201	--
Finland	--	0.080	0.307
France	0.468	0.280	0.224
Germany	0.160	0.170	0.288
Greece	0.151	0.196	0.288
Ireland	--	--	0.197
Italy	--	0.442	0.416
Japan	0.265	0.286	0.484
Netherlands	0.274	0.187	0.367
New Zealand	0.279	0.110	0.141
Norway	0.235	0.145	0.403
Portugal	--	0.390	0.427
South Korea	0.104	0.101	0.090
Spain	0.327	--	0.338
Sweden	--	--	--
Switzerland	0.489	0.362	0.396
United Kingdom	--	--	--
United States	0.294	0.169	0.146

Note: The table shows the ratio of M1 (currency held by the public plus checkable deposits) to nominal GDP. The data are from the International Monetary Fund, *International Financial Statistics*. Data on M1 were unavailable for Sweden and the United Kingdom.

I. The Demand for Money

We now extend the micro foundations of our model to consider the demand for money. Since we identify money with hand-to-hand currency, we assume that the interest rate paid on money is zero. In contrast, the rate of return on bonds and ownership of capital equals the interest rate, i, which we assume is greater than zero. Henceforth, we refer to bonds and ownership of capital as **interest-bearing assets**. The important point about these interest-bearing assets is that they yield a higher rate of return than money and are therefore better than money as long-term **stores of value**. Nevertheless, since households use money to make exchanges, households will hold some money for convenience, rather that always cashing in earning assets immediately prior to each exchange. That is, the demand for money will be greater than zero.

Households receive nominal labor income, wL, and nominal asset income, $i \cdot (B+PK)$, in the form of money. (We neglect profit because it is zero in equilibrium.) Households also use money to buy consumption goods, in the nominal amount PC, and to save, in the nominal amount $\Delta B + P \cdot \Delta K$. Although all these exchanges use money, it would be possible for households to hold little or no money at every point in time. If each inflow of income were perfectly synchronized with an equal outflow of expenditure on goods or purchases of interest-bearing assets, each household's money balance could always be zero. However, this synchronization would require a great deal of effort and planning. We assume, as a general matter, that households can reduce their average money balance by incurring more **transaction costs**. By transaction costs, we mean any expenses of time or goods related to the timing and form of various exchanges.

Examples of transaction costs are the time spent going to the bank or ATM machine and brokerage fees.

One way to maintain a low money balance is to rush off to the store as soon as wages are paid to spend one's entire weekly or monthly pay check on goods. Another method would be to go immediately to a financial institution to convert all of one's wage income into interest-bearing assets. More realistically, a household might immediately deposit its pay check into a checking or money-market account at a bank (or might arrange for the pay check to be deposited directly into an account). In addition, if workers were paid wages more frequently—say weekly rather than monthly—it would be easier for workers to maintain a lower average money balance.

The general idea is that, by putting more effort into money management and, thereby, incurring more transaction costs, households can reduce their average holding of money, M. For a given total of nominal assets, $M + B + PK$, a reduction in M raises the average holding of interest-bearing assets, $B + PK$. Since asset income is $i \cdot (B+PK)$, the rise in $B + PK$ raises asset income. Thus, a household's average holding of money, M, emerges from a tradeoff. With a frequent transaction strategy, M will be low and asset income will be high, but transaction costs will also be high. With an infrequent transactions strategy, M will be high and asset income will be low, but transaction costs will also be low. Households' choices of money holdings therefore involve finding the right balance between additional asset income and added transaction costs.

We use the term demand for money, labeled as M^d, to describe the average holding of money that results from households' optimal strategy for money management. Many formal models of money management have been developed to assess this demand for money. For our purposes, we do not have to go through this array of models. Rather, we are mainly interested in how some key variables in the model affect the quantity of money demanded, M^d. Specifically, we want to know how M^d depends on the price level, P, the interest rate, i, and real GDP, Y.

A. The price level and the demand for money

Suppose that the price level, P, doubles. Assume that the nominal wage rate, w, and the nominal rental price, R, also double, so that the real wage rate, w/P and the real rental price, R/P, do not change. In this case, household nominal income, $wL + i\cdot(B+PK)$, is twice as high as before.[3] However, household real income, $(w/P)\cdot L + i\cdot(B/P + K)$, is unchanged. Thus, we are considering a doubling of the nominal values of all variables, with no changes in the real values. In this circumstance, households would also want to double the nominal quantity of money, M, that they hold. This doubling of nominal money means that real money balances, M/P, do not change.

To think about this result, suppose that a household's nominal income is \$500 per week. Suppose that the initial plan for money management—involving some frequency of exchange between money and interest-bearing assets—dictates holding half a week's worth of income, on average, in the form of money. In this case, the household's average holding of money, M, would be \$250. After the doubling of the price level, P (along with the doubling of the nominal wage rate, w, and rental price, R), the household's nominal income is \$1000 per week. Households would still want to hold half a week's worth of income in the form of money. The reason that the number of weeks is unchanged is that

[3] We are assuming that L, K, and i are unchanged, and we are considering the average household for which B is zero.

the tradeoff for optimal money management—interest income versus transaction costs—remains the same. Hence, the household would not change the frequency with which it makes exchanges between money and interest-bearing assets. However, with nominal income twice as high, half a week's worth of income is twice as much money in nominal terms —\$500 instead of \$250. Therefore, the nominal demand for money, M^d, would double. Since M^d and P have both doubled, the ratio, M^d/P, is the same. That is, the **real demand for money**, M^d/P, is unchanged.

B. The interest rate and the demand for money

A higher interest rate, *i,* provides a greater incentive to hold down average holdings of money, *M*, in order to raise average holdings of interest- bearing assets, B+PK. That is, with a higher *i*, households would be more willing to incur transaction costs in order to reduce *M*. For example, they would respond to a higher *i* by transacting more frequently between money and interest-bearing assets. We predict accordingly that an increase in *i* reduces the nominal demand for money, M^d. For a given price level, *P*, we can also say that a higher *i* lowers the real demand for money, M^d/P.

C. Real GDP and the demand for money

Suppose again that a household holds half a week's worth of income in the form of money. Assume now that nominal income doubles—from \$500 per week to \$1000—while the price level, *P*, is unchanged. That is, real income, $(w/P)\cdot L + i\cdot(B/P + K)$, doubles. If its money-management plan were unchanged, each household would still hold half a week's worth of income as money. However, half a week's worth of income is now twice as much money—\$500 instead of \$250. Thus, households would double their nominal demand for money, M^d. Since P is constant, the real demand for money, M^d/P, also doubles.

This result has to be modified because higher real income shifts the tradeoff between interest income and transaction costs. Specifically, the larger real money balance, M/P, means that more real income on assets, $i \cdot (B/P + K)$, could be gained by spending additional effort on money management. As an example, when real income doubles, households may increase their frequency of transacting between money and interest-bearing assets. The greater transaction frequency means that, instead of holding 0.5 weeks' worth of income as money, households would hold only, say, 0.4 weeks' worth of income as money. In this case, nominal money demand, M^d, rises from \$250 to \$400, rather than \$500. That is, a doubling of real income raises M^d, but by less than 100%. Thus, the response of M^d is, in proportional terms, smaller than the change in real income. (This result is called **economies of scale in cash management**, because higher-income households hold less money in proportion to their income.) Since the price level, P, is unchanged, the real demand for money, M^d/P, rises, but less than proportionately, with real income.

In the aggregate, household real income moves along with real GDP, Y. To see why, recall that household real income is the sum of real wage income and real asset income:

$$real\ income = (w/P) \cdot L + i \cdot (B/P + K).$$

Since B is zero in the aggregate and $i = R/P - \delta$ (from equation [6.6]), we have

$$real\ income = (w/P) \cdot L + (R/P) \cdot K - \delta K.$$

We showed in chapter 6 that the real payments to the two factors, $(w/P)\cdot L$ to labor and $(R/P)\cdot K$ to capital, equaled real GDP, Y. Therefore, if we substitute Y for $(w/P)\cdot L$ + $(R/P)\cdot K$, we get

$$\begin{aligned} real\ income &= Y - \delta K \\ &= real\ net\ domestic\ product. \end{aligned}$$

Thus, for given depreciation, δK, the aggregate of household real income is determined by Y. We therefore have that the aggregate real demand for money, M^d/P, rises with Y.

D. Other influences on money demand

For given values of the price level, P, the interest rate, i, and real GDP, Y, money demand depends on the payments technology and the level of transaction costs. As examples, increased use of credit cards and greater convenience of checkable deposits reduce the demand for currency. Expanded use of ATM machines makes currency easier to obtain but has an uncertain impact on the demand for currency. The reason is that the machines make currency more attractive for payments but also make it easier to hold a smaller average currency balance by going more often to the ATM.

E. The money-demand function

We can summarize the discussion by writing down a formula for aggregate nominal money demand:

Key equation (money-demand function):

(10.1) $$M^d = P\cdot £(Y, i).$$

We assume in this equation that the nature of the transactions technology is given. Then, for given real GDP, Y, and the interest rate, i, nominal money demand, M^d, is proportional to the price level, P. For given P, M^d increases with Y (though less than proportionately) and falls with i. This dependence is summarized by the function $£(\cdot)$. Note that this function determines the real demand for money, M^d/P, that is, if we divide both sides of equation (10.1) by P, we get

$$M^d/P = £(Y, i). \qquad (10.2)$$

We call $£(\cdot)$ the real money-demand function. From this perspective, equation (10.1) says

nominal demand for money = price level times real demand for money.

The payments period and the demand for money

Irving Fisher (1971, pp. 83-85) stressed the dependence of the demand for money on the period between payments of wages. The general idea is that a shorter period makes it easier for workers to maintain a low average money balance. This effect is particularly important during extreme inflations—for example, during the German hyperinflation after World War I. In such situations, the cost of holding money becomes very high—we can represent this effect in our model by a high interest rate, i. Because of the high cost of holding money, workers and firms incurred more transaction costs—such as the costs of making more frequent wage payments—to reduce their average holdings of money. For 1923, the final year of the Germany hyperinflation, an observer reported, “it became the custom to make an advance of wages on Tuesday, the balance being paid on Friday. Later, some firms used to pay wages three times a week or even daily” (Bresciani-Turroni [1937, p. 303]). Similarly, during the Austrian hyperinflation after World War I, “The salaries of the state officials, which used to be issued at the end of the month, were paid to them during 1922 in installments three times per month” (J. van Walre de Bordes [1927, p. 163]).

Empirical evidence on the demand for money

Many statistical studies have analyzed the determinants of the demand for money. Most of these studies focus on M1, that is, the concept of money that comprises currency held by the public plus checkable deposits.

The empirical results confirm negative effects of interest rates on the demand for money, whether money is measured by M1 or currency. For example, in his classic empirical studies for the United States, Goldfeld (1973, 1976) found that a 10% increase in interest rates (for example, a rise from 5% to 5.5%) reduced the demand for M1 in the long run by about 3%. Goldfeld and Sichel (1990) and Fair (1987) reported similar findings for a number of OECD countries. Ochs and Rush (1983) showed that the negative effect of interest rates on the demand for M1 in the United States reflected similar proportionate effects on currency and checkable deposits. Mulligan and Sala-i-Martin (1996) showed that money demand became more sensitive to changes in interest rates when the level of interest rates rose. At low rates, say 2%, an increase in the interest rate by 10% (to 2.2%) lowered money demand by about 2%. However, at an interest rate of 6%, an increase in the interest rate by 10% (to 6.6%) reduced money demand by around 5%.

There is strong evidence for a positive effect of real GDP on real money demand and weaker evidence for economies of scale in this relation. Goldfeld (1973, 1976) found that an increase in real GDP by 10% led in the long run to an increase by about 7% in the real demand for M1. The change in M1 broke down into an increase of checkable deposits by around 6% and an increase in currency by about 10%. Therefore, economies of scale in the demand for M1 applied to checkable deposits but not to currency.

Our analysis predicted that an increase in the price level would raise the nominal demand for money in the same proportion. This proposition receives strong empirical support. For example, Goldfeld (1973, 1976) found that an increase in the price level by 10% led to a 10% increase in the nominal demand for M1.

We noted that changes in transactions technology can have important influences on the demand for money. This effect has been important in the

United States since the early 1970s because of a variety of financial innovations. The innovations included expanded use of credit cards, development of money-market accounts that were convenient alternatives to checkable deposits held at banks, introduction of ATM machines, and widespread use of electronic funds transfers.

Before the 1980s, economists ignored financial innovations when fitting equations for money demand. The estimated equations worked well up to the mid 1970s, in the sense that their predictions for the demand for money were fairly accurate. However, after the mid-1970s, the estimates that ignored financial innovations started to fail. In particular, the actual amount of M1—especially checkable deposits—that people held fell well below the amount predicted from earlier empirical relationships. Michael Dotsey (1985) showed that the volume of electronic funds transfers was a good measure of the extent of financial innovation. He found, first, that the spread of electronic funds transfers led to a substantial decline in the real demand for M1. Second, when the volume of electronic funds transfers was held constant, the demand for M1 became stable over time. In particular, the fitted equation showed effects from interest rates and real GDP that were similar to those, such as Goldfeld's, that ignored financial innovations but included data only up to the early 1970s.

II. Determination of the Price Level

We show in this section how to extend the model to determine the price level, P. The central idea is to add a new equilibrium condition: the quantity of money equals the quantity demanded.

A. The quantity of money equals the quantity demanded

We assume that money takes the form of currency. We assume further that the nominal quantity of money, M, is determined by the monetary authority, for example, the Federal Reserve in the United States or the European Central Bank in the euro zone. In the next chapter, we think of the monetary authority as setting the interest rate, i, rather than M.

The aggregate demand for nominal money is given by

(10.1) $$M^d = P \cdot \pounds(Y, i).$$

This equation gives the nominal quantity of money, M^d, that households want to hold, whereas M is the actual amount of nominal money outstanding. Thus, we propose as another equilibrium condition for our model the equality between the nominal quantity of money, M, and the nominal quantity demanded, M^d:

$$M = M^d.$$

If we substitute for M^d from equation (10.1), we get

(10.3) $$M = P \cdot \pounds(Y, i)$$

nominal quantity of money = nominal quantity demanded.

Recall that $\pounds(Y, i)$ is the function for the real demand for money, M^d/P. This real demand is determined, for a given transactions technology, by real GDP, Y, and the interest rate, i.

To see why we expect $M = M^d$ to hold, consider what happens when M differs from M^d. If M is greater than M^d, households have more money than they want to hold. Therefore, they try to spend their excess money, for example, on goods. This response affects the goods market. In particular, we anticipate that the increased desire to buy goods would raise the price level, P. Equation (10.3) says that this process continues until P rises enough to equate the right-hand side, which equals M^d, to the left-hand side, M. That is, the equilibrium price level is high enough so that households are willing to hold the outstanding quantity of money, M.

The same process works in reverse if M is less than M^d. In that case, households try to rebuild their money balances, for example, by reducing their spending on goods. In this case, P falls enough to equate M^d, given on the right-hand side of equation (10.3), to M.

One important point is that we are assuming that goods prices are flexible, that is, the price level, P, adjusts rapidly to ensure equality between the nominal quantity of money, M, and the nominal quantity demanded, M^d. This assumption about price flexibility parallels our previous assumptions about market-clearing conditions for the markets for labor and capital services. For the labor market, we assumed that the real wage rate, w/P, adjusted to ensure equality between the quantities of labor supplied, L^s, and demanded, L^d. For the rental market, we assumed that the real rental price, R/P, adjusted to ensure equality between the quantities of capital services supplied, $(\kappa K)^s$, and demanded, $(\kappa K)^d$. If we put all the assumptions together, we have that the three nominal prices—P, w, and R—adjust rapidly to ensure that three equilibrium conditions hold simultaneously: first, $M = M^d$; second, $L^s = L^d$; and third, $(\kappa K)^s = (\kappa K)^d$. Economists refer to this situation as one of **general equilibrium**.

Figure 10.1 shows graphically the equation of the nominal demand for money, M^d, to the nominal quantity of money, M. The vertical axis has the price level, P. The nominal demand for money, M^d, is the product of P and the real demand for money, $£(Y, i)$—see equation (10.1). Hence, for given $£(Y, i)$, M^d is proportional to P. We graph M^d accordingly as the upward-sloping red line, which starts from the origin. The nominal quantity of money, M, is shown by the vertical blue line.

The equilibrium condition is $M = M^d$, as shown in equation (10.3). Therefore, if the real demand for money, $£(Y, i)$, is given, the equilibrium price level is the value P^*, shown in Figure 10.1. At this point, the upward-sloping M^d line intersects the vertical M line. That is, P^* is the price level, P, that equates M^d to M.

B. A change in the quantity of money

To understand better how the model determines the price level, P, we can consider the effect from a one-time change in the nominal quantity of money, M. To be concrete, suppose that M doubles. The simplest way this could happen is if the monetary authority, on a one-time basis, printed up a lot of extra currency and gave it to people.

Figure 10.2 shows an increase in the nominal quantity of money from M to $2M$. Suppose, for the moment, that the money-demand line, M^d, did not shift. In this case, we can determine the change in the equilibrium price level, P^*, by looking at the intersections with the M^d line. The figure shows that the increase in nominal money from M to $2M$ raises the equilibrium price level from $P*$ to $2P*$. We can verify from equation (10.3) that the new equilibrium price level must be twice as high as the initial one. That is, if M doubles, P doubles, and real money balances, M/P, do not change.

Recall that the M^d line in Figure 10.2 is the product of the price level, P, and the real demand for money, $£(Y, i)$:

(10.1) $$M^d = P \cdot £(Y, i).$$

The upward slope of the M^d line in the figure reflects the proportional effect of P on M^d. Our assumption that the M^d line does not shift, when graphed versus P, amounts to the assumption that the real demand for money, $£(Y, i)$, does not change. In particular, we are assuming that real GDP, Y, and the interest rate, i, remain the same. We then found that a doubling of M led to a doubling of P and, hence, to no change in real money balances, M/P. This constancy of M/P is consistent with our assumption that the real demand for money, $£(Y, i)$, does not change. That is, after the doubling of M, we still have that real money balances, M/P, equal the real quantity demanded, $£(Y, i)$. Thus, our assumption that the M^d line does not shift in Figure 10.2 is satisfactory as long as Y and i do not change.

If we consider the labor market, described in Figure 8.3, we find that the market-clearing real wage rate, w/P, does not change. That is because our present analysis does not involve a shift in the technology level, A. Since w/P is unchanged, the quantity of labor input, L, is also unchanged. That is, L still equals the quantities demanded and supplied in the labor market. Since the price level, P, doubled, the constancy of w/P means that, in general equilibrium, the nominal wage rate, w, has to double.

Similarly, if we look at the rental market for capital services, described in Figure 8.7, we find that the market-clearing real rental price, *R/P*, does not change. Again, the reason is that the technology level, *A*, did not shift. Since *R/P* is unchanged, the quantity of capital services, κK, is also unchanged. That is, the utilization rate for capital, κ, and the capital stock, *K*, are the same as before. Since the price level, *P*, doubled and *R/P* is unchanged, in general equilibrium, the nominal rental price, *R*, has to double.

The interest rate, *i*, is determined by the real rental price, *R/P*, in accordance with our analysis from chapter 8:

interest rate = rate of return on ownership of capital

$$i = (R/P)\cdot\kappa - \delta(\kappa). \tag{10.4}$$

The expression on the right-hand side comes from equation (8.6). Since *R/P* and κ are unchanged, *i* is unchanged. That is, in general equilibrium, a one-time increase in nominal money, *M*, does not affect the interest rate.

Recall that real GDP, *Y*, is given from the production function:

$$Y = A\cdot F(\kappa K, L). \tag{8.3}$$

Since the technology level, *A*, the quantity of capital services, κK, and the quantity of labor, *L*, are all unchanged, *Y* is the same as before. In other words, in general equilibrium, a one-time increase in nominal money, *M*, does not affect real GDP.

We have verified that real GDP, *Y*, and the interest rate, *i*—the two determinants of real money demand, $£(Y, i)$—are unchanged. Therefore, $£(Y, i)$ does not change. This result validates our assumption in Figure 10.2 that the money-demand line, M^d, did not shift. We conclude that our previous result—that the price level, *P*, doubled—is correct.

To sum up, we find that a doubling of the nominal quantity of money, M, leads to a doubling of all of the nominal prices—the price level, P, the nominal wage rate, w, and the nominal rental price, R. Hence, there are no changes in real money balances, M/P, the real wage rate, w/P, and the real rental price, R/P. We also have that the determinants of the real demand for money, $\pounds(Y, i)$, remain the same. In particular, the increase in M has no effect on real GDP, Y, or the interest rate, i. Note, however, that nominal GDP equals PY. Since P doubles and Y is unchanged, nominal GDP doubles.

We can also show that the increase in the nominal quantity of money, M, does not change the division of real GDP, Y, between consumption, C, and gross investment, I. Consumption, C, does not change because its determinants—household real income, $(w/P)\cdot L + i\cdot(B/P + K)$, and the interest rate, i—are the same. Recall that real GDP, Y, must be divided between C and gross investment, I:

$$Y = C + I.$$

Since Y and C are the same, I must also be the same.

The analogous conclusions hold if we consider a decrease in the nominal quantity of money, M. For example, if M were to halve, P would halve, and real money balances, M/P, would again be unchanged. The nominal wage rate, w, and the nominal rental price, R, would also halve, so that the real wage rate, w/P, and the real rental price, R/P, would be the same. As before, the decrease in M has no effect on real GDP, Y, consumption, C, gross investment, I, and the interest rate, i. Hence, nominal GDP, PY, would fall to half its initial value.

C. The neutrality of money

The results exhibit a property called the **neutrality of money**. One-time changes in the nominal quantity of money, *M*, affect nominal variables but leave real variables unchanged. The real variables include real GDP, *Y*; consumption, *C*; gross investment, *I*; the real wage rate, *w/P*; the real rental price, *R/P*; and the quantity of real money balances, *M/P*. The interest rate, *i*, also does not change. We should think of *i* as a real variable because it governs intertemporal-substitution effects for consumption and work. In chapter 11, which introduces inflation, we distinguish the nominal interest rate from the real interest rate.

Almost all economists accept the neutrality of money as a valid long-run proposition. That is, in the long run, more or less nominal money, *M*, in the economy influences nominal variables but not real ones. However, many economists believe that the neutrality of money fails to hold in the short run. In particular, in the short run, increases in the nominal quantity of money, *M*, are usually thought to increase real GDP, *Y*, whereas decreases in *M* are thought to decrease *Y*. The main source of the difference in conclusions involves the flexibility of nominal prices, notably the price level, *P*, and the nominal wage rate, *w*. These nominal prices are thought to be flexible up or down in the long run in response to increases or decreases in *M*. However, *P* and *w* are often viewed as less flexible in the short run, especially when decreases in *M* mean that *P* and *w* have to decrease. In some models, the assumption of price flexibility is replaced by an assumption that *P* or *w* is sticky in the short run. We discuss sticky-price and sticky-wage models in chapter 16.

D. A change in the demand for money

We mentioned that financial innovations could shift the real demand for money. To explore the effects from a shift in the real demand for money, suppose that the nominal demand for money is given initially by

$$M^d = P \cdot £(Y, i),$$

where $£(Y, i)$ is the initial real demand for money. The nominal money demand, M^d, is graphed versus the price level, P, as the upward-sloping red line in Figure 10.3.

Suppose now that a change in the transactions technology raises the real demand for money to $[£(Y, i)]'$, so that the nominal demand becomes

$$(M^d)' = P \cdot [£(Y, i)]'.$$

We graph the new nominal demand, $(M^d)'$, as the upward-sloping green line in Figure 10.3. Note that, at any price level, P, the nominal quantity of money demanded is higher along the green line than along the red line.

We assume that the nominal quantity of money is fixed at M, shown by the vertical blue line in Figure 10.3. Therefore, the initial equilibrium price level is the value P^*, where M^d equals M. After the rise in the real demand for money, the equilibrium price level is the value $(P^*)'$, where $(M^d)'$ equals M. Note that the rise in the real demand for money leads to a lower price level, that is, $(P^*)'$ is lower than P^*. (As before, we assume that the price level adjusts rapidly to its equilibrium level.)

The increase in the real demand for money is similar to a reduction in the nominal quantity of money, M, in that the price level, P, falls in each case. However, one difference between the two cases is that a change in M is fully neutral, whereas a change in the real demand for money is not fully neutral. To see why, note that the increase in the real demand for money led to a fall in P for a given M. Therefore, the real quantity of money, M/P, increased. In addition, the change in transactions technology that led to the increase in the real demand for money—perhaps an expanded use of ATM machines—would itself have real effects. For example, the resources used up in transaction costs would change. However, in most cases, these effects will be small enough to neglect.

E. The cyclical behavior of the price level and real money balances: theory

In chapters 7-9, we studied how shifts to the technology level, A, generated economic fluctuations. In chapter 7, increases in A raised real GDP, Y, and also increased consumption, C, gross investment, I, the real wage rate, w/P, the real rental price of capital, R/P, and the interest rate, i. In chapter 8, increases in A also raised labor input, L, and the utilization rate for capital, κ. In chapter 9, expansions of L showed up partly as rises in employment rates, as well as increases in average hours worked and the labor force. Now we can use our analysis of the demand for money to determine how changes in A affect the price level, P.

Recall that the nominal demand for money is given by

(10.1) $$M^d = P \cdot \pounds(Y, i).$$

In an economic expansion, the rise in real GDP, Y, raises the real demand for money, $\pounds(Y, i)$, on the right-hand side. However, we also found that the interest rate, i, tends to increase in an economic expansion. The increase in i tends to reduce $£(Y, i)$. The overall change depends on the magnitudes of the increases in Y and i and on the sensitivity of $£(Y, i)$ to Y and i. Typical estimates indicate that $£(Y, i)$ increases overall. One reason is that the rise in i tends to be relatively small. Another reason is that the sensitivity of $£(Y, i)$ to i tends to be relatively low. Therefore, we assume that the real demand for money, $£(Y, i)$, increases overall during an economic expansion.

We can use Figure 10.3 to determine the effect of an economic expansion on the price level, P. Recall that this figure applied to an increase in the real demand for money, $£(Y, i)$, caused by a change in the transactions technology. However, the same construction applies if $£(Y, i)$ increases for some other reason. In the present case, $£(Y, i)$ rises because of the changes in real GDP, Y, and the interest rate, i. Thus, we can use the figure to study the effect of an economic expansion on the price level, P.

We see from Figure 10.3 that, for a given nominal quantity of money, M, the increase in the real demand for money, $£(Y, i)$, lowers the price level, P. Hence, a relatively low P tends to accompany the increase in real GDP, Y. Thus, our model has a new prediction: if M, does not vary, P will be countercyclical. That is, P will be relatively low in booms and relatively high in recessions.[4]

Before we evaluate this prediction, we should note a second prediction of the model. We know that the nominal quantity of money, M, equals the nominal quantity demanded, $M^d = P \cdot £(Y, i)$, that is,

(10.3) $$M = P \cdot £(Y, i).$$

Therefore, if we divide both sides by P, we get

(10.5) $$M/P = £(Y, i)$$

[4] This result was first emphasized by Kydland and Prescott (1990).

real quantity of money = real quantity demanded.

We have assumed that the right-hand side, *£(Y, i)*, increases in a boom. Therefore, we also have on the left-hand side that *M/P* increases in a boom. In other words, *M/P* is procyclical—high in a boom and low in a recession.

F. The cyclical behavior of the price level and real money balances: empirical

Now we consider how the model's predictions about prices and real money balances match up with the U.S. data. We measure the price level, *P*, by the deflator for the gross domestic product. We use our usual procedure to construct the cyclical part of the log of *P*. The result is the blue graph in Figure 10.4. The red graph reproduces the cyclical part of the log of real GDP, from Figure 7.2. We can see from Figure 10.4 that the price level typically fluctuates in the direction opposite to real GDP.[5] That is, as predicted, the price level is countercyclical. The results are similar if we use the consumer price index (CPI), rather than the GDP deflator, to measure the price level.[6]

We see from the blue graph in Figure 10.4 that two of the largest positive deviations of the GDP deflator from its trend were in 1974-75 and 1981-82. These increases in the price level reflected, in part, the sharp rises in oil prices generated by the oil cartel run by OPEC (the Organization of Petroleum Exporting Countries). These

[5] From 1954.1 to 2004.3, the correlation of the cyclical part of the GDP deflator with the cyclical part of real GDP is -0.63. The GDP deflator is less variable than real GDP: the standard deviation of the cyclical part of the GDP deflator is 0.8%, compared with 1.6% for real GDP.

[6] From 1954.1 to 2004.3, the correlation of the cyclical part of the CPI with the cyclical part of real GDP is -0.56. The standard deviation of the cyclical part of the CPI is 1.2%. Davis and Kanago (2004) argue that, from the standpoint of checking the macroeconomic model, it is better to consider co-movements in matched measures of output and prices—such as real GDP and the GDP deflator—rather than unmatched ones—such as real GDP and the CPI. The reason is that unmatched sets are affected by changes in relative prices, for example, movements in prices of goods contained in GDP but not in the market basket used to calculate the CPI.

periods also featured recessions in the United States, so that real GDP was well below trend, as seen on the red graph. Thus, the price level and real GDP moved in opposite directions at the times of these oil shocks. However, the inverse relation between the price level and real GDP is not just the result of oil shocks—the pattern applies more generally to the period 1954-2004 shown in Figure 10.4.[7]

To consider real money holdings, we start by measuring nominal money, *M*, by currency held by the public.[8] Real money balances, *M/P*, are then the ratio of currency to the GDP deflator. The blue graph in Figure 10.5 shows the cyclical part of the log of real currency. The figure shows that real currency tended to move in the same direction as real GDP.[9] That is, as predicted, real currency is procyclical. We can also see from the figure that real currency was less variable than GDP from the 1950s through the mid 1980s. However, real currency became more variable than GDP since the late1980s. This change may reflect increased volatility of the demand for U.S. currency in foreign countries.

We can also consider the cyclical behavior of other real monetary aggregates, such as those based on M1 and M2. The U.S. data on M1 and M2, as currently defined by the Federal Reserve, are available only since 1959. Over the period 1959.1 to 2004.3, real M1 and real M2 are both procyclical. The correlation of the cyclical part of real GDP with the cyclical part of real M2 is about the same as that with the cyclical part of real currency. However, the correlation of the cyclical part of real GDP with the cyclical

[7] The pattern is different before World War II. For example, from 1880 to 1940 (using the available annual data), the correlation of the cyclical part of the GNP deflator with the cyclical part of real GNP is weakly positive, 0.27. The main difference in the earlier period seems to be the importance of banking panics, which tended to reduce both the price level and output. This effect was especially important during the Great Depression, from 1929 to 1933.
[8] This concept excludes currency held in the vaults of financial institutions.
[9] From 1954.1 to 2004.3, the correlation of the cyclical part of real currency with the cyclical part of real GDP is 0.49. The standard deviation of the cyclical part of real currency is 1.2%.

part of real M1 is substantially lower.[10] In addition, real M1 has become highly variable since the mid-1980s. The likely reason is the development of new forms of deposits, such as convenient money-market funds, which are close substitutes for checkable deposits but are not included in M1.

To sum up, we find that the empirical evidence accords with the two new predictions from the theory. First, the price level is countercyclical. Second, various measures of real money balances are procyclical.

G. Price-level targeting and endogenous money

Up to now, we have assumed that the nominal quantity of money, M, is constant. This assumption is unrealistic. First, the various measures of M have positive trends, that is, the average growth rate of money is greater than zero. Second, each measure of M tends to fluctuate around its trend. Some of these fluctuations reflect random variations, not controlled by the monetary authority. However, much of the variation likely reflects purposeful monetary policy, aimed at achieving desired values of some of the macroeconomic variables.

We assume here that the monetary authority can determine the path of the nominal quantity of money, M, possibly subject to some uncontrollable random errors. The assumption that the monetary authority can control M is reasonable if we take a narrow view of money as currency held by the public.[11] However, the assumption that

[10] From 1959.1 to 2004.3, the correlation of the cyclical part of real GDP is 0.52 with the cyclical part of real M2 and 0.33 with the cyclical part of real M1. The standard deviations of the cyclical parts are 1.9% for real M2 and 2.3% for real M1.

[11] Another reasonable assumption is that the monetary authority can control a slightly broader nominal aggregate, the **monetary base**, which adds the reserves of financial institutions (held as vault cash or as deposits at the central bank) to the currency held by the public.

the monetary authority can control the nominal quantity of money becomes less obvious if we take a broader view of M, for example, to add the many types of deposit accounts included in M2. For our main analysis, we think of M in a narrow sense, as the currency held by the public.

We assume now that the monetary authority's objective is to adjust the nominal quantity of money, M, to achieve stability of the price level, P. This objective is called **price-level targeting**. In this analysis, M becomes an endogenous variable. That is, M is determined by the model to be the value consistent with the monetary authority's objective, in this case, maintenance of a stable P.

The key relation is still the equality between the nominal quantity of money, M, and the nominal demand for money, $P \cdot £(Y, i)$:

$$M = P \cdot £(Y, i). \tag{10.3}$$

Price-level targeting implies that the monetary authority chooses M at each point in time to maintain a constant value of P. We see from equation (10.3) that P can be constant only if M varies to compensate for changes in the real demand for money, $£(Y, i)$. For example, if $£(Y, i)$ doubles but M less than doubles, P will fall. Alternatively, if $£(Y, i)$ doubles but M more than doubles, P will rise. To keep P fixed, the proportionate change in M has to equal the proportionate change in $£(Y, i)$. This condition will tell us how price-level targeting determines the trend and cyclical characteristics of M.

To consider the trend in the nominal quantity of money, M, we should return to the analysis of long-run economic growth from the Solow model of chapter 5. We consider the long-run or steady-state situation, in which real GDP, Y, grows at a constant rate due to technological progress and population growth. The interest rate, i, is constant in the steady state. Therefore, the growth of *Y* produces a continuing rise in the real demand for money, *£(Y, i)*. If we think of money as currency, the empirical estimates of money demand suggest that the growth rate of the real demand for money will be about the same as the growth rate of *Y*. [12] In this case, maintenance of a constant price level, *P*, requires *M* to grow at the same rate as *Y*. Thereby, the growth rate of *M* matches the growth rate of *£(Y, i)* and allows *P* to remain constant. That is, in equation (10.3), the growth rate of *M* on the left-hand side equals the growth rate of *£(Y, i)* on the right-hand side. In chapter 11, we look in detail at long-run growth rates of *M*.

Now we consider the cyclical behavior of the nominal quantity of money, *M*. The important result from before is that the real demand for money, *£(Y, i)*, is high in a boom and low in a recession. We also found that, if *M* did not vary, the price level, *P*, would fall in a boom and rise in a recession. That is, *P* would be countercyclical. If the monetary authority wants to avoid this countercyclical movement of *P*, it has to adjust *M* to match the cyclical variations in the real demand for money, *£(Y, i)*. We can see this result from the equality between *M* and the nominal demand for money:

$$M = P \cdot £(Y, i). \tag{10.3}$$

Since *£(Y, i)* rises in a boom, *M* has to increase proportionately by the same amount to avoid a fall in *P*. Similarly, since *£(Y, i)* falls in a recession, *M* has to decrease proportionately by the same amount to avoid a rise in *P*.

We have found that price-level targeting dictates that the nominal quantity of money, *M*, should be relatively high in booms and relatively low in recessions. In other

[12] With economies-of-scale in money demand, the real demand for money would grow at a slower rate than real GDP. As discussed before, the empirical evidence suggests that these economies of scale are important for checkable deposits but not for currency. We are also neglecting the possibility that continuing financial innovations would affect the real demand for money.

words, *M* should be procyclical. Recall, however, that we already found that the price level, *P*, is countercyclical. Therefore, the monetary authority (the Federal Reserve) apparently has not pursued a monetary policy that is sufficiently procyclical to eliminate the countercyclical behavior of *P*. Nevertheless, we would like to know whether the authority has followed a policy that is somewhat procyclical, that is, whether *M* is above trend during booms and below trend during recessions. Such a policy would have moderated the countercyclical pattern for *P*.

Empirically, nominal money, *M*, does not have a strong cyclical pattern. If we measure *M* by currency held by the public, the cyclical part of *M* is only weakly positively associated with the cyclical part of real GDP: the correlation from 1954.1 to 2004.3 is 0.10. Figure 10.6 shows the patterns for the cyclical parts of nominal currency and real GDP. Similar results hold for broader definitions of money.[13]

Our predictions for the cyclical behavior of real money holdings, *M/P*, do not depend on whether the monetary authority adjusts nominal money, *M*, to target the price level, *P*. The cyclical pattern in *M/P* has to reflect the cyclical movements in the real demand for money, *£(Y, i)*. We have already assumed that the cyclical changes in real GDP, *Y*, and the interest rate, *i*, lead overall to procyclical variations in *£(Y, i)*. That is, the real demand for money is high in a boom and low in a recession. The important point is that this procyclical pattern in *M/P* will be the same for any cyclical pattern in *M*.[14] In this sense, our earlier prediction that *M/P* would be procyclical still accords with the data.

---

[13] From 1959.1 to 2004.3, the correlation of the cyclical part of nominal M1 with the cyclical part of real GDP is 0.16. The correlation of the cyclical part of nominal money with the cyclical part of real GDP rises to 0.33 if we define money to be M2. However, M2 includes various deposits, such as money-market funds, that are not readily controlled by the monetary authority.

[14] This conclusion holds as long as the variations in nominal money, *M*, are neutral. In this case, different patterns in *M* do not affect the patterns in real GDP, *Y*, and the interest rate, *i*, and, hence, do not affect the cyclical behavior of the real demand for money, *£(Y, i)*.

Recall our finding that real currency holdings and other measures of real money balances are procyclical.

Seasonal variations in money

We have argued that, to achieve price-level stability, the monetary authority would want to vary the nominal quantity of money, M, to match the changes in the real demand for money, $£(Y, i)$, that occur during economic fluctuations. A similar argument applies to the regular variations in $£(Y, i)$ associated with the seasons. For example, until the mid 1980s, the quantity of real currency held in December was about 2% higher than the average for the year, whereas the amount held in February was about 1% lower than average. If the monetary authority kept the nominal quantity of currency constant over the year, the price level, P, would tend to have the reverse seasonal pattern—low in December and high in February. We get this result from the equality between M and the nominal demand for money, $P\cdot£(Y, i)$, in equation (10.3). If M were fixed, P would be low when $£(Y, i)$ was high (December) and high when $£(Y, i)$ was low (February).

In practice, the Federal Reserve has adjusted the nominal quantity of currency to match the seasonal variations in the real demand for currency.

Therefore, the price level, P, did not have any significant tendency to vary with the seasons. For example, up to the mid 1980s, nominal currency tended to be about 2% higher than average in December and about 1% lower than average in February.

The seasonal variations in the real demand for U.S. currency have declined substantially since the mid 1980s. For example, the December excess of real currency over the average for the year varied between 1.6% and 2.2% from 1950 to 1983 but then fell to 1.0% in 1990, 0.8% in 2000, and 0.5% in 2003. A study by the Federal Reserve and the U.S. Treasury (Board of Governors of the Federal Reserve System [2003]) suggests that this change relates to the increased use of U.S. currency in foreign countries. The foreign demand for U.S. currency does not have the same seasonal pattern as that found in domestic demand. Therefore, the full seasonal variation became weaker when more of the currency was held by foreigners.

The pattern is different if we look at checkable deposits, the other part of M1. The seasonal variations in the U.S. real demand for checkable deposits have not changed so much over time. For example, the December excess of real checkable deposits over the average for the year was between 2.4% and 3.3% from 1959 to 1996 and then rose to 3.6% in 2003. The difference from currency can probably be explained from the observation that the foreign demand for U.S.

checkable deposits is not nearly as important as the foreign demand for U.S. currency.

III. Summing Up

We extended our macroeconomic model to add another equilibrium condition: the nominal quantity of money, M, equals the nominal quantity demanded, M^d. We then have a general-equilibrium model that determines the price level, P, the nominal wage rate, w, and the nominal rental price, R. The three nominal prices are flexible and adjust rapidly to ensure that three equilibrium conditions hold: $M = M^d$, $L^s = L^d$ (clearing of the labor market), and $(\kappa K)^s = (\kappa K)^d$ (clearing of the market for capital services).

We extended the microeconomic foundations of the model to consider the determinants of the nominal demand for money, $M^d = P \cdot £(Y, i)$, where the function $£(Y, i)$ gives the quantity of money demanded in real terms, M^d/P. The real quantity demanded, M^d/P, rises with real GDP, Y, and falls with the interest rate, i. Shifts in the financial technology also affect M^d/P.

An increase in the nominal quantity of money, M, raises nominal variables—such as P, w, and R—in the same proportion. Real variables—such as w/P, R/P, and real GDP, Y—do not change. This property is called neutrality of money. An increase in the real quantity of money demanded, M^d/P, reduces P if M is fixed. This result suggests that P would be counter-cyclical. However, if the monetary authority stabilizes P, M adjusts endogenously to accommodate changes in M^d/P. In this case, M would be procyclical.

Questions and Problems

Mainly for review

10.1. What are the costs of transacting between money and alternative financial assets? You might want to make a list and include such items as the time spent going to the bank or waiting in line. How were these costs affected by the development of automatic teller machines (ATMs)?

10.2. Consider the following changes and state whether the effect on the real demand for money is an increase, decrease, or no change:

a. an increase in the nominal interest rate, i.

b. an increase in real transaction costs.

c. an increase in real GDP, Y, caused by a rise in per capita real GDP with population held constant.

d. an increase in real GDP, Y, caused by a rise in population with per capita real GDP held constant.

e. an increase in the price level, P.

10.3. Suppose that the nominal quantity of money, M, doubles once and for all.

a. The rise in the price level, P, suggests that workers will be worse off. Is this correct?

b. The rise in the nominal wage rate, w, suggests that workers will be better off. Is this right?

c. How do your results relate to the concept of the neutrality of money?

10.4. Some economists, called monetarists, believe that changes in the price level, *P*, are primarily the result of changes in the quantity of money, *M*. Can this conclusion be based solely on theoretical reasoning?

10.5. Explain why a favorable shock to the production function tends to reduce the price level, *P*. Why might the reaction of monetary policy prevent the fall in *P*?

10.6. Explain why it is important to distinguish between shifts in the nominal quantity of money, *M*, and shifts in the demand for money, M^d. What association would we expect between the price level, *P*, and real GDP, *Y*, for periods in which both types of monetary shifts occurred?

10.7. What is the meaning of the term "endogenous money?" Under what circumstances would endogenous money generate a positive association between nominal money, *M*, and real GDP, *Y*?

Problems for discussion

10.x. Effects of other variables on the demand for money

Assume given values of real GDP, Y, population, the nominal interest rate, i, and real transaction costs. If these variables are given, would you say that the following statements about the real demand for money are true, false, or uncertain?

a. Agricultural societies have lower real money demand than industrial societies.

b. Dictatorships have higher real money demand than democracies.

c. Countries with a larger fraction of persons who are elderly have higher real money demand.

d. Countries with a higher literacy rate have lower real money demand.

For empirical evidence on these effects, see the study by Lawrence Kenny (1991).

10.x. Transaction frequency and the demand for money

Suppose that a household's consumption expenditure is \$60,000 per year and is financed by monthly withdrawals from a savings account.

a. Show on a graph the pattern of the household's money holding over a year. What is the average money balance? Should we identify this average balance with the demand for money in our model?

b. Suppose now that the frequency of withdrawals from the savings account rises to two per month. What happens to the average money balance?

c. Return to part a. but assume now that consumption expenditure is \$120,000 per year. If withdrawals of the savings account are made monthly, what is the average money balance? How does this average compare to the one in part a.? Is it optimal for the frequency of withdrawals to remain the same when consumption expenditure increases? Explain.

10.x. Velocity of money

The velocity of money is the ratio of the dollar volume of transactions, say nominal GDP, divided by the nominal quantity of money. How is the velocity of money affected by

a. an increase in the nominal interest rate, i?

b. an increase in real GDP, Y caused by a rise in per capita real GDP with population held constant.

c. an increase in real GDP, Y caused by a rise in population with per capita real GDP held constant.

d. an increase in the price level, P?

e. Why might nominal GDP not be the correct measure of transactions?

f. What do you predict happens to the velocity of money as an economy develops?

10.x. The payments period and the demand for money

Suppose that a worker has an annual income of $60,000. Assume that the worker receives wage payments twice per month. The worker keeps all of his or her wages in money, does not use any alternative financial assets, and pays for consumption expenditure of $60,000 per year from money holdings.

a. What is the worker's average money balance?

b. What would the average money balance be if the worker were paid monthly, rather than twice per month?

c. What is the general relation between the payments period and the demand for money?

d. How do the results change if the worker puts part of his or her monthly wages into a savings account and then makes withdrawals as needed from this account?

10.x. Shopping trips and the demand for money

Assume again the conditions in the first part of question 10.x, with a worker paid once per month. However, instead of making consumption expenditure as a uniform flow, the worker (or the worker's spouse) makes periodic shopping trips. At each trip, enough goods (for example, groceries) are bought to last until the next trip.

a. If the worker shops four times per month, what is the average money balance? Why is the answer different from that in part a. of question 10.x?

b. What happens if the worker shops only twice per month?

c. What is the general relation between the frequency of shopping trips and the demand for money?

d. Suppose that the cost of a shopping trip rises, perhaps because of an increase in the price of gasoline. What would happen to the frequency of shopping trips? What happens to the demand for money?

10.x. Transaction costs and households' budget constraints

In our model, we neglected the resources that households use up in transaction costs. Suppose that these costs take the form of purchases of goods and services (such as fees paid to banks or brokers). Assume that real transaction costs decline because of the expansion of ATM machines.

a. How does this change show up in households' budget constraints?

b. What is the income effect on consumption and leisure?

c. Suppose that transaction costs represent the time required to go to a bank, rather than a purchase of goods and services. Is there a change in the results in parts **a.** and **b.**?

10.x. A currency reform

Suppose that the government replaces the existing monetary unit with a new one. For example, the United States might shift from the old dollar to the Reagan dollar, defined to equal 10 old dollars. People would be able to exchange their old currency for the new one at the ratio of 10 to 1. Also, any contracts that were written in terms of old dollars are reinterpreted in Reagan dollars at the ratio 10 to 1.

a. What happens to the price level, P, and the interest rate, i?

b. What happens to real GDP, Y, consumption, C, and labor, L?

c. Do the results exhibit the neutrality of money?

10.x. Denominations of currency

Consider how people divide their holdings of currency between large bills (say $100 bills) versus small ones. How would the dollar fraction of currency held in large bills depend on the following:

a. the price level, P?

b. real per capita GDP?

c. population?

d. incentives to avoid records of payments, for example, for tax evasion or to disguise illegal activities, such as the drug trade?

e. increased holdings of U.S. currency in foreign countries?

Given these results, the factors for the United States are not so easy to explain. The fraction of dollar value of currency (including coins) held in denominations of $100 and higher stayed nearly constant—between 20% and 22%—from 1944 to 1970. Then the fraction rose steadily to reach 69% in 2004. What do you think explains these patterns?

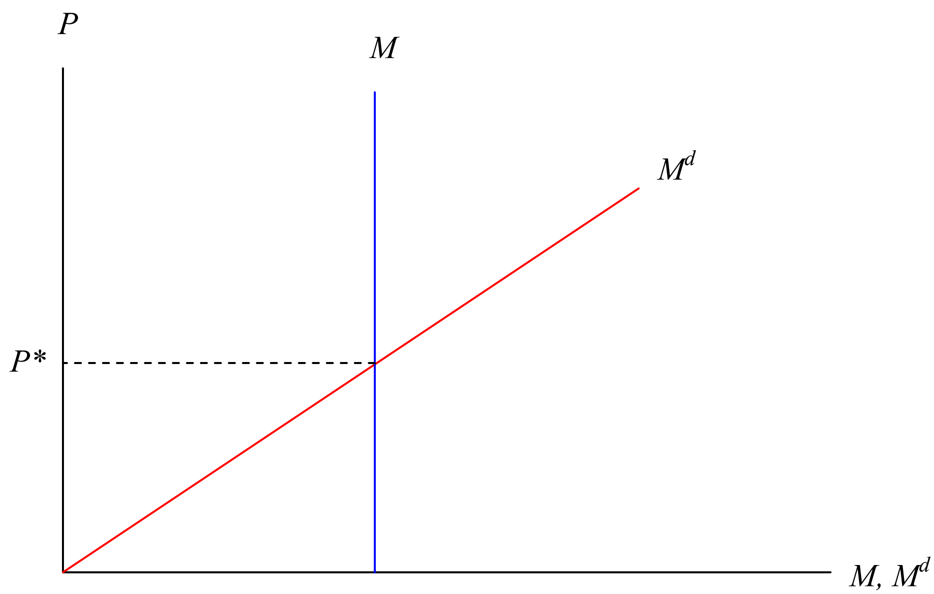


Figure 10.1

The Demand for Money Equals the Quantity of Money

The nominal demand for money is $M^d = P\cdot £(Y, i)$ from equation (10.1). For a given real demand for money, $£(Y, i)$, M^d is proportional to the price level, P. Therefore, M^d is given by the upward-sloping red line, which starts from the origin. The nominal quantity of money is the constant M, shown by the vertical blue line. The condition $M = M^d$ holds when the price level is P^*. Thus, P^* is the equilibrium value of P.

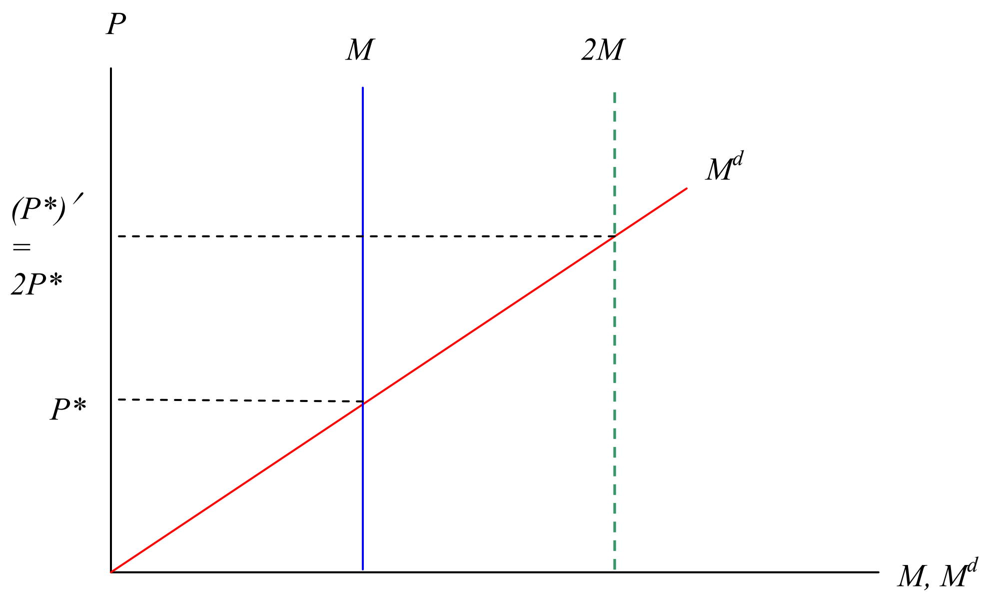


Figure 10.2

An Increase in the Quantity of Money

The nominal demand for money, M^d, shown by the red line, is the same as in Figure 10.1. The nominal quantity of money doubles from *M* (the vertical blue line) to *2M* (the vertical green line). Therefore, the equilibrium price level doubles, on the vertical axis, from *P** to *(P*)′ = 2P**.

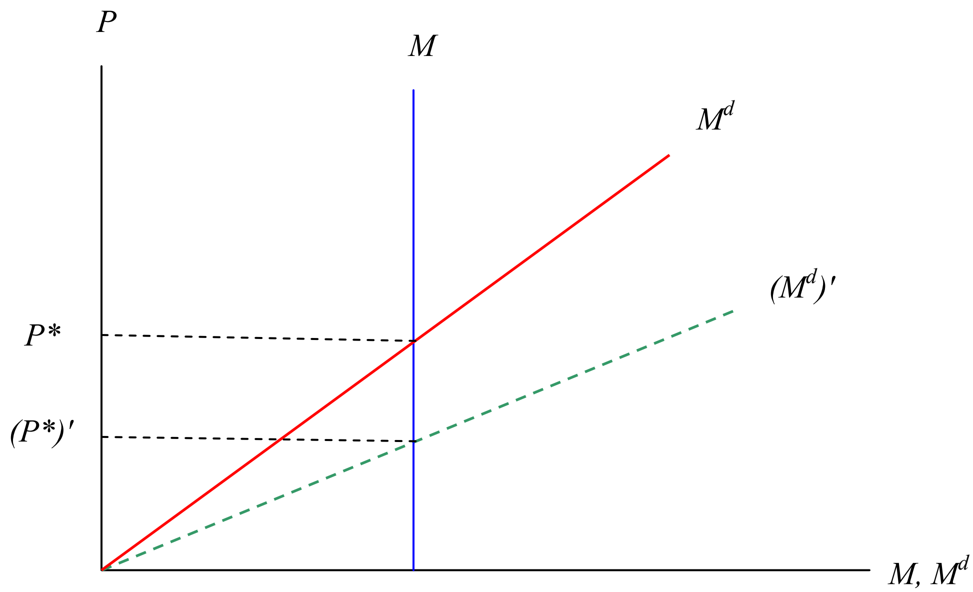


Figure 10.3

An Increase in the Demand for Money

The nominal demand for money is $M^d = P \cdot ￡(Y, i)$ from equation (10.1). We consider an increase in the real demand for money, $￡(Y, i)$. This real demand is higher along the dashed green line, $(M^d)'$, than along the red line, M^d. The nominal quantity of money is the constant M, shown by the vertical blue line. The increase in the real demand for money lowers the equilibrium price level from P^* to $(P^*)'$.

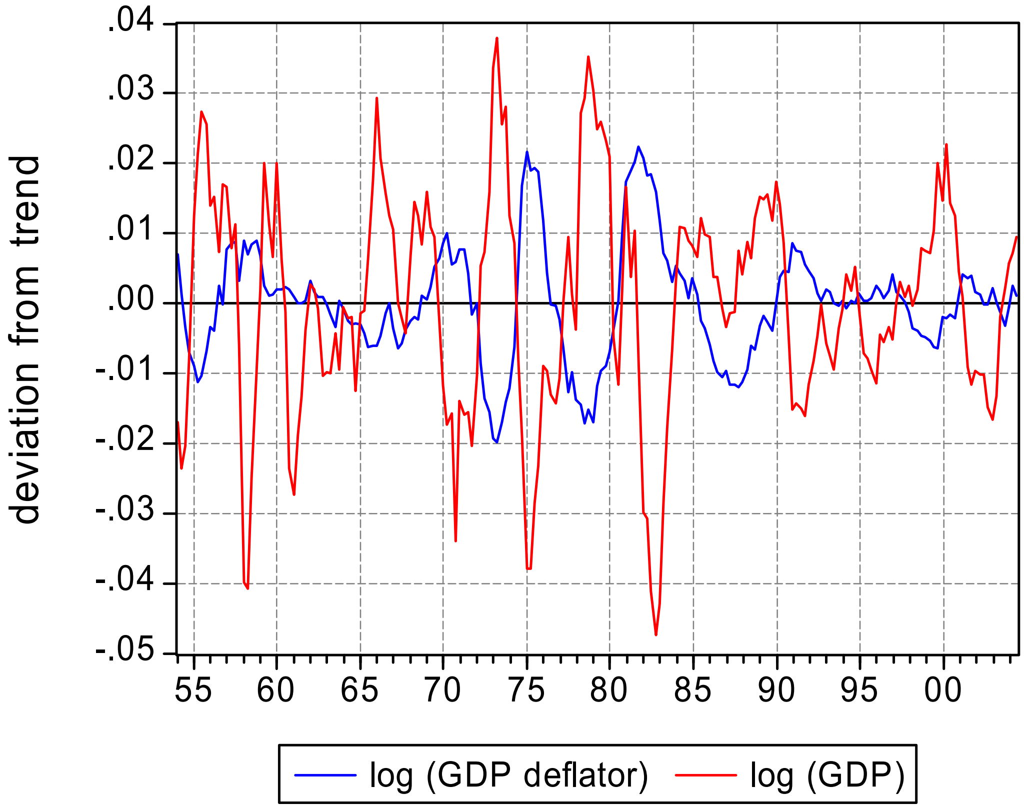


Figure 10.4

Cyclical Parts of U.S. GDP and the Price Level

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of the GDP deflator from its trend. The GDP deflator is countercyclical—it fluctuates in the direction opposite to real GDP—and is less variable than real GDP.

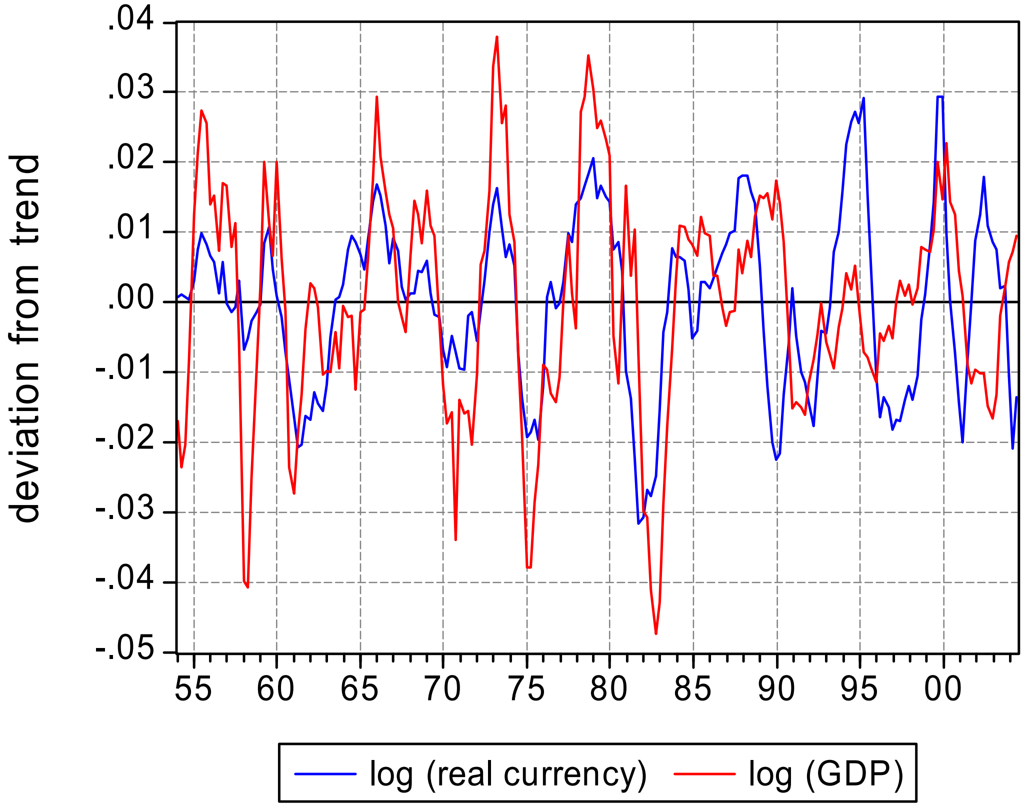


Figure 10.5

Cyclical Parts of U.S. GDP and Real Currency

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of real currency from its trend. Real currency is the ratio of nominal currency held by the public to the GDP deflator. Real currency is procyclical—it fluctuates in the same direction as real GDP. For the whole sample, 1954.1-2004.3, real currency holdings were less variable than real GDP. However, real currency became more variable than real GDP since the late 1980s.

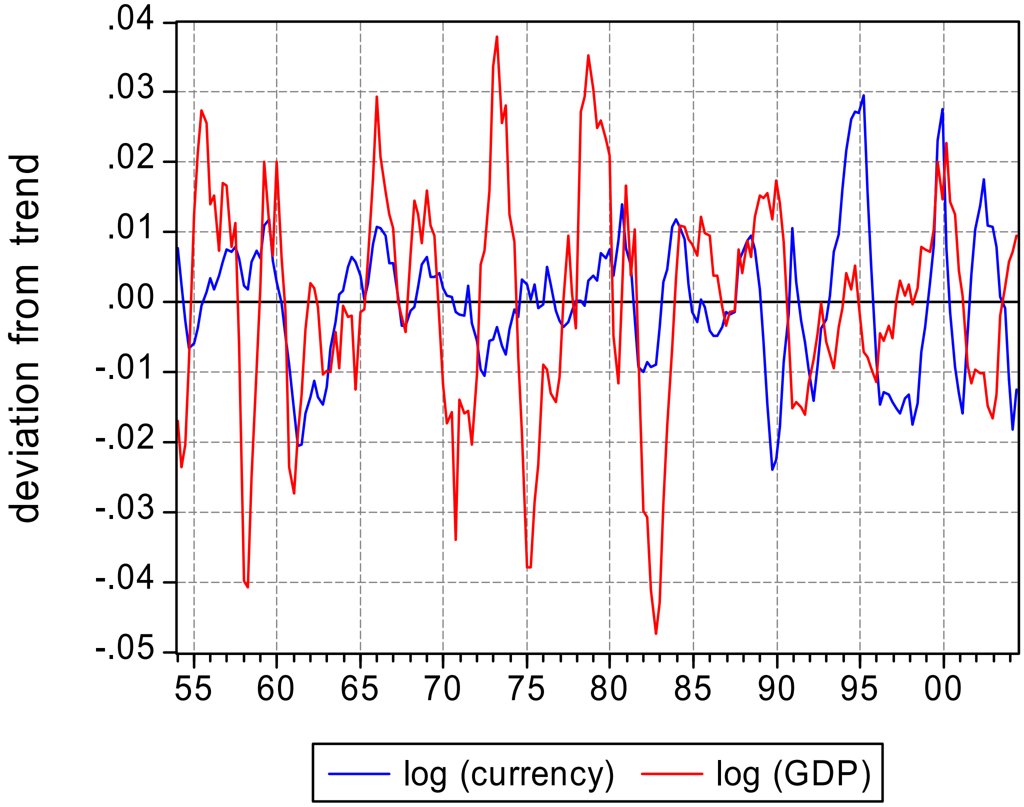


Figure 10.6

Cyclical Parts of U.S. GDP and Nominal Currency

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of nominal currency held by the public from its trend. Nominal currency is weakly procyclical. Nominal currency was less variable than real GDP over the full sample, 1954.1-2004.3. However, nominal currency became more variable than real GDP since the late 1980s.

January 12, 2005

Chapter 11

Inflation, Money Growth, and Interest Rates

In chapter 10, we considered the determination of the price level, *P*. Now we consider **inflation**, by which we mean a continuing upward movement in *P*. Our previous analysis suggests possible sources of inflation. To sort out the possibilities, we should think about the condition that all money be willingly held:

(10.3) $$M = P \cdot £(Y, i),$$

nominal quantity of money = nominal quantity demanded.

We found in chapter 10 that one way for the price level, *P*, to increase is through a downward shift in the real demand for money, $£(Y, i)$. This change could reflect a financial innovation, such as increased use of credit cards and electronic funds transfers. The downward movement in the quantity of real money demanded could also reflect a reduction in real GDP, *Y*, due to a decrease in the technology level, *A*. Notice, however, each decrease in real money demand creates a single increase in *P*, rather than a continuing series of increases in *P*. To generate inflation along these lines, we would need a succession of reductions in the real demand for money. Although this pattern is theoretically possible, this possibility is not borne out by the data. The typical pattern for most countries is that long-run economic growth steadily raises real GDP, *Y*, and thereby continually expands the real demand for money, $£(Y, i)$. Although these effects could be offset by financial innovations that reduced $£(Y, i)$, the usual pattern is that, on net, the real quantity of money, *M/P*, trends upward. Thus, the real quantity of money demanded must also be trending upward. Consequently, *P* would trend downward if the nominal quantity of money, *M*, did not change. Thus, we cannot explain sustained inflation from regular downward movements in the real demand for money.

The remaining suggestion from the previous analysis is a link between inflation and increases in the nominal quantity of money, *M*. Empirically, it is clear that *M*—measured as currency or as broader aggregates such as M1—typically grows over time. Moreover, **money growth rates** differ substantially across countries and over time for a single country. Therefore, variations in money growth rates are good candidates for explaining inflation. To assess this linkage, we begin by considering data on inflation and money growth across countries.

I. Cross-Country Data on Inflation and Money Growth

Table 11.1 shows inflation rates and money growth rates for 82 countries that have data from 1960 to 2000.[1] We measure the price level, *P*, for each country by the consumer price index. The inflation rate, shown in column 1, is the growth rate of *P* from 1960 to 2000. The table lists the countries in descending order of the inflation rate. Column 2 shows the growth rate of nominal money, *M*, defined as currency. Column 3 shows the difference between the growth rate of currency and the inflation rate. This measure tells us the growth rate of real currency, that is, of *M/P*. Column 4 has the growth rate of real GDP. Finally, column 5 shows the inflation rate for a more recent period, 1980-2000.

---

[1] Because of missing data, some of the starting dates are a few years after 1960 and some of the ending dates are a few years before 2000.

Table 11.1

Inflation Rates and Money Growth Rates for 82 Countries

Country	1960-2000 Inflation rate (1)	Growth rate of currency (2)	Growth rate of real currency (3)	Growth rate of real GDP (4)	1980-2000 Inflation rate (5)
Congo (Brazz.)	0.831	0.820	-0.011	-0.005	1.456
Brazil	0.818	0.844	0.026	0.049	1.292
Argentina	0.700	0.695	-0.005	0.025	0.911
Peru	0.535	0.528	-0.007	0.033	0.893
Uruguay	0.407	0.390	-0.017	0.019	0.388
Bolivia	0.371	0.391	0.020	0.026	0.629
Chile	0.343	0.402	0.059	0.041	0.137
Turkey	0.313	0.340	0.027	0.045	0.471
Israel	0.285	0.323	0.038	0.055	0.370
Ghana	0.269	0.263	-0.006	0.033	0.293
Sierra Leone	0.248	0.227	-0.020	0.016	0.453
Indonesia	0.228	0.303	0.075	0.060	0.104
Mexico	0.212	0.251	0.039	0.045	0.335
Ecuador	0.195	0.244	0.050	0.040	0.314
Iceland	0.186	0.194	0.008	0.039	0.173
Tanzania	0.182	0.196	0.015	0.029	0.224
Colombia	0.174	0.212	0.038	0.042	0.199
Venezuela	0.165	0.180	0.015	0.024	0.284
Nigeria	0.157	0.190	0.033	0.019	0.223
Jamaica	0.143	0.177	0.034	0.019	0.180
Iran	0.130	0.179	0.049	0.047	0.192
Costa Rica	0.125	0.165	0.039	0.043	0.188
Paraguay	0.119	0.168	0.048	0.044	0.161
Madagascar	0.117	0.131	0.014	0.018	0.158
Algeria	0.110	0.155	0.045	0.040	0.128
Kenya	0.110	0.138	0.028	0.041	0.128
Haiti	0.107	0.122	0.015	0.048	0.126
Dominican Rep.	0.106	0.144	0.037	0.052	0.153
Greece	0.104	0.141	0.037	0.037	0.131
Portugal	0.103	0.115	0.012	0.042	0.103
Philippines	0.099	0.133	0.034	0.039	0.102
Burundi	0.096	0.103	0.007	0.017	0.106
Syria	0.093	0.150	0.056	0.058	0.128
South Korea	0.091	0.177	0.086	0.075	0.055
Egypt	0.091	0.129	0.038	0.049	0.121
El Salvador	0.090	0.095	0.004	0.030	0.127
Rwanda	0.089	0.097	0.009	0.033	0.088
South Africa	0.088	0.117	0.030	0.033	0.111
Gambia	0.086	0.120	0.034	0.043	0.101

Honduras	0.086	0.122	0.036	0.035	0.121
Guatemala	0.085	0.119	0.035	0.039	0.120
Spain	0.082	0.122	0.040	0.041	0.064
Nepal	0.081	0.148	0.067	0.040	0.090
Mauritius	0.080	0.116	0.035	0.044	0.072
Sri Lanka	0.080	0.116	0.036	0.039	0.104
Trinidad	0.078	0.091	0.013	0.034	0.079
India	0.077	0.116	0.039	0.048	0.086
Italy	0.076	0.100	0.023	0.030	0.064
Pakistan	0.076	0.115	0.039	0.056	0.078
Cameroon	0.073	0.086	0.013	0.035	0.064
Fiji	0.071	0.100	0.029	0.034	0.052
Barbados	0.070	0.096	0.026	0.043	0.041
Jordan	0.070	0.092	0.022	0.052	0.052
New Zealand	0.069	0.063	-0.005	0.024	0.060
Ireland	0.068	0.101	0.033	0.048	0.050
Ivory Coast	0.066	0.093	0.027	0.039	0.054
United Kingdom	0.065	0.061	-0.005	0.024	0.047
Gabon	0.062	0.097	0.035	0.049	0.047
Senegal	0.061	0.085	0.025	0.027	0.048
Togo	0.060	0.103	0.043	0.018	0.050
Ethiopia	0.060	0.083	0.023	0.029	0.055
Finland	0.060	0.088	0.029	0.033	0.042
Denmark	0.057	0.065	0.008	0.026	0.039
Sweden	0.056	0.065	0.009	0.025	0.048
Australia	0.056	0.086	0.031	0.037	0.050
Norway	0.055	0.061	0.006	0.035	0.048
Morocco	0.053	0.103	0.050	0.049	0.054
France	0.052	0.052	0.000	0.033	0.039
Niger	0.051	0.063	0.011	0.013	0.030
Thailand	0.051	0.105	0.055	0.067	0.044
Canada	0.045	0.075	0.029	0.037	0.039
Cyprus	0.044	0.098	0.053	0.054	0.046
United States	0.044	0.074	0.030	0.035	0.037
Japan	0.043	0.101	0.058	0.050	0.014
Belgium	0.041	0.037	-0.004	0.031	0.032
Netherlands	0.040	0.052	0.012	0.032	0.024
Luxembourg	0.039	0.093	0.054	0.042	0.032
Austria	0.038	0.058	0.020	0.033	0.029
Switzerland	0.034	0.041	0.007	0.022	0.026
Malaysia	0.033	0.087	0.054	0.065	0.033
Germany	0.032	0.065	0.033	0.022	0.024
Singapore	0.032	0.098	0.066	0.081	0.023

Notes: Countries are listed in descending order of the inflation rate for 1960-2000. Countries included are those with data from at least the early 1960s through the late 1990s. In some cases, the starting date is a few years after 1960 and the ending date is a few years before 2000. The inflation rate is based on consumer price indexes. Data are from International Monetary Fund, *International Financial Statistics*.

Note the following from the table:

- The inflation rate was greater than zero for all countries from 1960 to 2000 (column 1) and from 1980 to 2000 (column 5). That is, all countries had some degree of inflation. Falling prices—called **deflation**—did not apply to any country, at least in terms of the overall experience since 1960 or 1980. The lowest inflation rate was 1.4 percent per year for Japan from 1980 to 2000.[2]
- The growth rate of nominal currency (column 2) was also greater than zero for all countries from 1960 to 2000. (The same result holds from 1980 to 2000, although these data are not shown in the table.)
- The median inflation rate for the 82 countries from 1960 to 2000 was 8.3 percent per year, with 30 of them exceeding 10%. For the growth rate of nominal currency, the median was 11.6% per year, with 50 above 10%.
- There is a broad cross-sectional range for the inflation rates and the growth rates of money. The inflation rates varied from 83% for Congo (Brazzaville) and 82% for Brazil to 3.2% for Singapore and Germany. The growth rates of currency varied from 84% for Brazil and 82% for Congo (Brazzaville) to 3.7% for Belgium and 4.1% for Switzerland.
- In most countries, the growth rate of nominal currency, *M*, exceeded the growth rate of prices, *P* (the inflation rate). Therefore, the growth rate of real money balances, *M/P*, shown in column 3, was greater than zero in most countries. The median growth rate of real currency from 1960 to 2000 was 3.0% per year. We discuss in section III.F how this growth related to the growth rate of real GDP.

[2] From 1995 to 2000, the inflation rate for Japan was only 0.3 percent per year.

- Column 5 of the table shows that some countries that had high inflation rates from 1960 to 2000 managed to attain lower inflation rates in the more recent period, 1980 to 2000. For example, Chile went from an inflation rate of 34 percent per year from 1960 to 2000 (56 percent per year from 1960 to 1980) to 14 percent per year from 1980 to 2000. Similarly, Indonesia went from an inflation rate of 23 percent per year from 1960 to 2000 (34 percent per year from 1960 to 1980) to 10 percent per year from 1980 to 2000. More recently, some other previously high-inflation countries reduced their inflation rates. For example, from 1995 to 2000, the inflation rates were only 6.2 percent per year in Israel, 2.7 percent per year in Iceland, 6.6 percent per year in Peru, and -0.1 percent per year in Argentina (which, however, returned to high inflation in 2001). More generally, however, there is a tendency for a country that has a high inflation rate in one period to have a high inflation rate in another period. For example, for the 82 countries shown in the table, the correlation of the inflation rate from 1960 to 1980 with that from 1980 to 2000 was 0.58. The corresponding correlation for growth rates of nominal currency was 0.54.
- For the purpose of understanding inflation, the most important observation from the cross-country data is the remarkably strong positive association between the inflation rate and the growth rate of nominal currency. Figure 11.1 displays this relationship for the cross-country data from 1960 to 2000. The vertical axis has the inflation rate (column 1 of the table), and the horizontal axis has the growth rate of nominal currency (column 2). The graph makes clear that a country that had a lot of inflation also had a high money growth rate. The correlation between

the two variables is extremely high, 0.99. The slope is close to one—that is, an increase in the growth rate of nominal currency by 1% per year is associated with an increase in the inflation rate by around 1% per year. This strong association does not isolate the direction of causation between inflation and money growth. That is, we cannot say whether a country had a high inflation rate because it had a high money growth rate or vice versa. However, we can be sure that a country cannot have a high inflation rate over 40 years unless it also has a high money growth rate.

One important suggestion from the cross-country data is that, to understand inflation, we have to include money growth as a central part of the analysis. That is, we have to take seriously Milton Friedman's famous observation that "inflation is always and everywhere a monetary phenomenon."[3] Thus, we now add the various pieces needed to extend our model to allow for inflation and money growth.

II. Inflation and Interest Rates

We now begin to incorporate inflation into the model. We start by considering actual and expected inflation rates.

A. Actual and expected inflation

Suppose that the price level in year 1 is P_1, whereas that the next year is P_2. The change in the price level from year 1 to year 2 is $\Delta P_1 = P_2 - P_1$. Let π denote the

[3]See Friedman (1968b, p. 29).

inflation rate. The inflation rate from year 1 to year 2, which we label as π_1, is the ratio of the change in the price level to the initial price level:

$$\pi_1 = (P_2 - P_1)/P_1 \tag{11.1}$$

$$= \Delta P_1/P_1.$$

For example, if $P_1 = 100$ and $P_2 = 105$, the inflation rate from year 1 to year 2 is

$$\pi_1 = (105 - 100)/100$$

$$= 0.05 \text{ per year (or 5\% per year)}.$$

Table 11.1 showed that inflation rates, π, are typically greater than zero. Therefore, we usually consider cases where prices are rising, so that $P_2 > P_1$ and, hence, $\pi_1 > 0$. We can, however, also study situations in which prices are falling, so that $P_2 < P_1$ and, hence, $\pi_1 < 0$. These cases are called deflations. Economists have recently become more interested in the topic of deflation because the experiences of a few countries, notably Japan, suggest that deflations may become empirically relevant in the future.

We can rearrange equation (11.1) to solve out for year 2's price level, P_2. First, multiply through by P_1 to get

$$\pi_1 \cdot P_1 = P_2 - P_1.$$

Then add P_1 to both sides, combine the two terms that have P_1 and swap the left- and right-hand sides to get

$$P_2 = (1 + \pi_1) \cdot P_1. \tag{11.2}$$

Thus, over one year, the price level rises by the factor $1 + \pi_1$. For example, if $P_1 = 100$ and $\pi_1 = 0.05$,

$$P_2 = (1.05) \cdot 100 = 105.$$

In making decisions, such as the choice between consuming this year or next year, households want to know how prices will change over time. Since the future is unknown, households have to form forecasts or **expectations of inflation**. We use the symbol π_1^e to denote the expectation of the inflation rate π_1. Usually we think of households as forming this expectation during year 1. If they already know year 1's price level, P_1, the prediction for year 1's inflation rate, π_1, corresponds to a forecast of next year's price level, P_2. That is, from equation (11.1), the forecast for P_2 will determine the forecast of π_1.

Since future price levels are unknown, forecasts of inflation will be imperfect. Therefore, the actual inflation rate, π_1, will typically differ from its expectation, π_1^e. Hence, the forecast error—or **unexpected inflation**—is usually non-zero. Sometimes unexpected inflation is greater than zero and sometimes it is less than zero. Although these errors are unavoidable, households are motivated to keep the errors as small as possible. Therefore, they will use the available information on past inflation and other variables to avoid systematic mistakes. Expectations formed this way are called **rational expectations**.[4] This rationality implies that unexpected inflation would not exhibit a systematic pattern of errors over time. For example, if unexpected inflation is greater than zero this year, it may be either greater or less than zero next year.

B. Real and nominal interest rates

Let i_1 be the interest rate on bonds in year 1. Suppose, for example, that $i_1 = 5\%$ *per year* $= 0.05$ *per year*. If a household holds $1000 of bonds in year 1, how

[4] The idea of rational expectations comes from Muth (1961). For a discussion of applications to macroeconomics, see Lucas (1977).

much would the household have in year 2? First, the household still has the principal of \$1000. Then, second, the household has interest income equal to \$1000*0.05 = \$50. Therefore, the total is

$$\textit{principal (\$1000) + interest (\$1000*0.05) = total assets in year 2 (\$1000+\$1000*0.05)}$$

$$= \$1000*(1 + 0.05)$$

$$= \$1000 + \$50$$

$$=\$1050.$$

Now we generalize to allow for any interest rate, i_1. The principal carried over to year 2 is still \$1000. The interest income equals \$1000* i_1. Therefore, we have

$$\textit{principal (\$1000) + interest (\$1000*}\ i_1\textit{) = total assets in year 2 (\$1000+\$1000*}\ i_1\textit{)}$$

$$= \$1000*(1+ i_1).$$

Thus, the dollar value of assets held as bonds rises over the year by the factor $1+ i_1$. We can think of i_1 as the dollar or **nominal interest rate**. The important point is that the nominal interest rate determines the change over time in the nominal value of assets held as bonds.

This analysis is fine for seeing how a household's nominal assets change over time. However, households do not really care about the nominal value of their assets. What they really care about is the goods they can buy with these assets, that is, with the real value of assets. Thus, we have to figure out what happens over time to the real value of assets held as bonds.

If the price level, P, is constant, as in previous chapters, the change in the nominal value of assets held as bonds also determines the change in the real value of assets held as bonds. For example, suppose that the price level in years 1 and 2 is

$$P_1 = P_2 = 100.$$

If we again assume $i_1 = 0.05$ per year, \$1000 of bonds in year 1 becomes \$1050 of bonds in year 2. The real values of these holdings are

$$\textit{year 1:}\ \ 1000/100 = 10$$

and

$$\textit{year 2:}\ \ 1050/100 = 10.5.$$

Thus, year 2's real value exceeds year 1's real value by 5%:

$$(10.5 - 10)/10 = 0.5/10 = 0.05.$$

Since the real value of assets held as bonds grows at 5% per year, we can say that the **real interest rate** is 5% per year. By real interest rate, we mean the rate at which the real value of assets held as bonds changes over time. The key point is that, when the price level is constant, the real interest rate equals the nominal interest rate.

However, the real and nominal interest rates will not be the same if the inflation rate, π_1, is greater or less than zero. As an example, suppose that $\pi_1 = 1\%$ *per year* $= 0.01$ per year, as shown in row 1 of Table 11.2. Equation (11.2) shows that the price

Table 11.2

Nominal and Real Interest Rates

		year 1	**year 2**
(1)	**inflation rate**	0.01	
(2)	**price level**	100	101
(3)	**nominal interest rate**	0.05	
(4)	**nominal assets**	1000	1050
(5)	**real assets**	10	10.4
(6)	**real interest rate**	0.04	

level rises over one year by the factor $1+\pi_1$. Thus, when $\pi_1 = 0.01$, the price level increases from $P_1 = 100$ in year 1 to $P_2 = 100*(1.01) = 101$ in year 2. These values are in row 2 of the table.

Suppose again that the nominal interest rate is $i_1 = 0.05$ per year, as shown in row 3 of Table 11.2. Thus, nominal assets still grow from \$1000 in year 1 to \$1050 in year 2, as shown in row 4 of the table.

The question is what happens to real assets? In year 1, real assets are

$$\textit{year 1 real assets} = 1000/100 = 10 \textit{ goods}.$$

In year 2, real assets are

$$\textit{year 2 real assets} = 1050/101 = 10.4 \textit{ goods}.$$

These values are in row 5 of Table 11.2. We can see that real assets have increased over the year by the proportion

$$(10.4 - 10)/10$$

$$= 0.4/10$$

$$= 0.04.$$

We have defined the real interest rate to be the rate at which the real value of assets held as bonds changes over time. Therefore, as shown in row 6, the real interest rate in this example is 0.04 per year = 4% per year. Notice that the real interest rate, 4%, falls short of the nominal interest rate, 5%, by the inflation rate, 1%.

Now we want to generalize to allow for any nominal interest rate, i_1, and inflation rate, π_1. Since the nominal interest rate is i_1, dollar assets rise over the year by the factor $1+ i_1$:

$$dollar\ assets\ in\ year\ 2 = (dollar\ assets\ in\ year\ 1)*(1+ i_1).$$

Since the inflation rate is π_1, the price level rises over the year by the factor $1+\pi_1$ (see equation [11.2]):

$$P_2 = P_1 \cdot (1+\pi_1).$$

If we divide the first equation by the second one, we get

$$\frac{dollar\ assets\ in\ year\ 2}{P_2} = \left(\frac{dollar\ assets\ in\ year\ 1}{P_1}\right) \bullet \frac{(1+i_1)}{(1+\pi_1)}.$$

Since the ratio of dollar assets to the price level gives real assets, we have

$$real\ assets\ in\ year\ 2 = (real\ assets\ in\ year\ 1) \cdot (1+i_1)/(1+\pi_1).$$

In other words, real assets rise over the year by the factor $(1+i_1)/(1+\pi_1)$.

We define the real interest rate, denoted by r_1, to be the rate per year at which assets held as bonds change in real value. Given this definition, real assets would rise over one year by the factor $1+ r_1$. Therefore, we must have

(11.3) $$(1+r_1) = (1+i_1)/(1+\pi_1).$$

For the example in Table 11.2, we had

$$1.04 \approx 1.05/1.01,$$

so that the real interest rate was $r_1 \approx 0.04$.

In the general case, we can get a useful formula for the real interest rate, r_1, if we manipulate equation (11.3). Multiply through on both sides by $1+\pi_1$ to get

$$(1+r_1) \cdot (1+\pi_1) = 1+i_1.$$

If we multiply out the two terms on the left-hand side, we get

$$1 + r_1 + \pi_1 + r_1 \cdot \pi_1 = 1 + i_1.$$

If we simplify by canceling out the "1" on each side and placing all terms except for r_1 on the right-hand side, we have

$$r_1 = i_1 - \pi_1 - r_1 \cdot \pi_1.$$

The right-hand side has the cross term, $r_1 \cdot \pi_1$. This term tends to be small; for example, if $r_1 = 0.04$ and $\pi_1 = 0.01$, the term is 0.0004. Moreover, it turns out that this cross term appears in the formula for r_1 only because we allowed interest rates and inflation rates to be compounded only once per year. A more accurate procedure would be to compound these rates continuously. In that case, the cross term would not appear at all. Hence, we would get the simpler (and more accurate) formula for the real interest rate:

Key equation:

real interest rate = nominal interest rate minus inflation rate

(11.4) $$r_1 = i_1 - \pi_1.$$

We shall use equation (11.4) for our subsequent comparisons of real and nominal interest rates.

C. The real interest rate and intertemporal substitution

We discussed intertemporal-substitution effects on consumption in chapter 6. We noted that a higher interest rate, i_1, motivated households to reduce year 1's consumption, C_1, compared to year 2's, C_2. However, when the inflation rate, π_1, is not zero, it is the real interest rate, r_1, rather than the nominal rate, i_1, that matters for intertemporal substitution. To see why, suppose that a household reduces C_1 by one unit and thereby raises real assets held as bonds or capital by one unit. These extra real assets will become $1+r_1$ additional real assets in year 2. Therefore, the household can raise C_2 by $1+r_1$ units.

Thus, one unit less of C_1 can be transformed into $1+r_1$ units more of C_2. If r_1 rises, the incentive to defer consumption increases. Therefore, the household would react by reducing C_1 and raising C_2.

We have found that the real interest rate, r_1, has an intertemporal-substitution effect on consumptions, C_1 and C_2. In contrast, if we hold fixed r_1, a change in the nominal interest rate, i_1, does not have this effect. For example, if the inflation rate, π_1, rises by 1% per year and i_1 also rises by 1% per year—so that $r_1 = i_1 - \pi_1$ does not change—households would have no incentive to shift between C_1 and C_2. That is because the terms on which C_1 can be exchanged for C_2 depend on r_1, which has not changed.

The same conclusions apply to intertemporal-substitution effects on labor supply. Again, it is the real interest rate, r_1, rather than the nominal rate, i_1, that matters. If r_1 rises, households have more incentive to defer leisure. Therefore, the household would react by reducing year 1's leisure and raising year 2's leisure. Or, to put it another way, the household would raise year 1's labor supply, L_1^s, and reduce year 2's, L_2^s.

D. Actual and expected real interest rates

We usually think of financial instruments, such as U.S. Treasury bills, that specify in advance the nominal interest rate, i. For example, a newly issued 3-month U.S. Treasury bill (called a T-bill) guarantees the nominal interest rate when held for three months. The real interest rate on this kind of nominal bond depends on the inflation rate over the three months.

As an example, during year t, the real interest rate on a 3-month T-bill is

(11.5) $$r_t = i_t - \pi_t.$$

We can think of i_t as the nominal interest rate on a 3-month T-bill issued on January 1 of year *t*. The rate i_t would be expressed at an annual rate, such as 0.02 per year. The variable π_t is the inflation rate, also expressed at an annual rate, from January to April. That is, the inflation rate applies to the three months covered by the T bill. The problem is that this inflation rate is unknown in January when a household decides to buy the T-bill. The real interest rate, r_t, becomes known only later, when π_t can be measured.

We already mentioned that households form expectations of inflation, π^e. Suppose that, in January, households expect the inflation rate from January to April to be π_t^e. This expected inflation rate determines the **expected real interest rate**, r_t^e, on the T-Bill from equation (11.5):

(11.6) $$r_t^e = i_t - \pi_t^e$$

expected real interest rate = nominal interest rate minus expected inflation rate.

For example, if $i_t = 0.02$ per year and $\pi_t^e = 0.01$ per year, $r_t^e = 0.01$ per year.

We mentioned that the actual inflation rate, π_t, might be greater than or less than its expectation, π_t^e. If the inflation rate turns out to be surprisingly high, so that $\pi_t > \pi_t^e$, equations (11.5) and (11.6) tell us that the real interest rate, r_t, will be less than its expectation, r_t^e. The opposite applies when the inflation rate turns out to be surprisingly low. In other words, if we combine equations (11.5) and (11.6), we get

$$r_t - r_t^e = (i_t - \pi_t) - (i_t - \pi_t^e).$$

Therefore, if we cancel out i_t and $-i_t$, we get

(11.7) $$r_t - r_t^e = -(\pi_t - \pi_t^e).$$

Thus, errors in forecasts of inflation, π_t - π_t^e, create errors of the opposite sign in forecasts of real interest rates, r_t - r_t^e.

When households decide about consumption and labor supply, they know the expected real interest rate, r_t^e, rather than the actual rate, r_t. Thus, intertemporal-substitution effects will depend on r_t^e—because decisions cannot be based on a variable, r_t, that will not be known until later. For this reason, we would like to be able to measure r_t^e. To do so, we need a measure of the expected inflation rate, π_t^e.

1. Measuring expected inflation. Economists have used three methods to measure the expected inflation rate:

1. Ask a sample of people about their expectations.

2. Use the hypothesis of rational expectations, which says that people's expectations correspond to optimal forecasts, given the available information. Then use statistical techniques to measure these optimal forecasts.

3. Use market data, including various interest rates, to infer expectations of inflation.

The main shortcoming of the first approach is that the sample may not be representative of the whole economy. Also, economists have a better theory of how people take actions than of how they answer questions on surveys. Nevertheless, the survey approach can be useful, and we discuss its application to expected inflation in the next section.

The second approach, based on rational expectations, has produced some successes and some problems. One challenge is to figure out what information people have when they form expectations. Another issue is the choice among statistical models to generate forecasts of inflation. In any event, the measures of expected inflation from the rational-expectations approach turn out to conform reasonably well to those from the survey findings, which we discuss in the next section.

The third approach, which relies on market data, has become increasingly applicable since the governments of many advanced countries began in the 1980s and 1990s to issue **indexed bonds**. Unlike nominal bonds, which specify in advance the nominal interest rate, indexed bonds guarantee in advance the real interest rate. For example, a 10-year indexed bond adjusts its nominal payouts of interest and principal to ensure a specified real rate of return over the 10 years. We discuss in a later section how to use the data on these bonds to infer expected inflation rates, π_t^e.

2. U.S. expected inflation and interest rates since World War II. A commonly used survey measure of expected inflation is the one initiated in 1946 by Joseph Livingston, a Philadelphia journalist. The survey asks about 50 economists for their forecasts of the consumer price index (CPI) 6 and 12 months in the future.[5] These forecasts of future price levels allow us to construct expected inflation rates, π_t^e. The 6-month-ahead forecasts of inflation are shown as the red graph in Figure 11.2. The figure also shows as the blue graph the actual inflation rate, π_t, over the previous 12 months. These values of π_t would have been available to the survey respondents at the time they were making their forecasts.

[5] For a discussion of the Livingston survey, see Carlson (1977).

Figure 11.2 shows that actual and expected inflation rates tended to move together from 1948 to 2004. Inflation rates were low from the mid 1950s to the mid 1960s, rose until the start of the 1980s, then fell sharply in the early 1980s. Inflation rates have been low and fairly stable since the mid 1980s. For example, in mid 2004, the expected inflation rate was below 2%.

Actual inflation rose above expectations in the late 1960s and again at the times of the oil shocks in 1973-75 and 1979-81. Actual inflation fell below expectations in the early and mid 1980s, when inflation rates came down sharply.

Figure 11.3 shows as the blue graph the nominal interest rate, i_t, on 3-month U.S. Treasury bills. The red graph is the expected real interest rate, r_t^e. This rate is calculated by subtracting the Livingston measure of the expected inflation rate, π_t^e, shown in Figure 11.2, from i_t.[6] That is, we use the formula

(11.6) $$r_t^e = i_t - \pi_t^e .$$

The nominal interest rate, i_t, moved upward from the 1950s to the early 1980s. However, because the expected inflation rate, π_t^e, rose in a similar way, the expected real interest rate, r_t^e, did not have this upward trend. This tendency for i_t and π_t^e to move together is the typical long-run pattern. Therefore, we will want to see in our theory why this pattern would apply.

The expected real interest rate, r_t^e, was fairly stable at 2-3% from the mid 1950s until the early 1970s. Then r_t^e fell to near zero for much of the 1970s, before rising to around 4% in the 1980s. The rate r_t^e was again near zero in 1992-93, rose back to 3-4%

[6] One problem, likely minor, is that the time horizons do not exactly match. The T-Bill rate is for three months, whereas the Livingston Survey expectation is for six months.

for the rest of the 1990s, then fell sharply to near zero from mid 2001 through 2004. We shall also want to use our theory to understand the determination of this expected real interest rate.

3. Indexed bonds, real interest rates, and expected inflation rates. Much more reliable measures of real interest rates and expected inflation rates are now available from data on indexed government bonds. These bonds adjust the nominal payouts of interest and principal in response to changes in consumer-price indexes. Thereby, the bonds guarantee a pre-specified real interest rate, r_t, up to the maturity of each issue. The U.K. government first issued these types of bonds in 1981. Subsequently, indexed bonds were issued by the governments of Australia (1985), Canada (1991), Iceland (1992), New Zealand (1995), Israel (1995), the United States (1997), Sweden (1997), and France (1998).[7]

The market prices of indexed bonds allow us to calculate the real interest rate, r_t, for bonds of various maturities. The red graph in Figure 11.4 shows the data since 1997 for U.S. indexed bonds of 10-year maturity. Data since 1998 are shown for 30-year maturity (the green graph) and since 2001 for 5-year maturity (the blue graph). The real interest rates, r_t, from 1997 to 2000 on the 10-year bonds ranged from 3.3% to 4.3% These rates, which peaked early in 2000, are slightly higher than the expected real interest rates, r_t^e, shown for this period for 3-month U.S. Treasury bills in Figure 11.3. Figure 11.4 shows that, following the high point in 2000, the real rates, r_t, on the indexed

[7] The data on real yields are available, for example, from Global Financial Data, Inc.

bonds plunged. In 2003-04, the rate on the 10-year issue was around 2%, that on the 5-year issue was close to 1%, and that on the 30-year issue was a little over 2%.

Figure 11.5 shows the real interest rates, r_t, in three other advanced countries that have issued indexed bonds for some time: the United Kingdom, Canada, and Australia. The U.K. data, available since 1983, show that r_t was typically between 3% and 4% from 1983 through 1997. These rates were similar to those in the United States from 1997 to 2000. However, the U.K. rate decreased earlier than the U.S. one—the U.K. rate fell in 1998 to around 2% and remained around that level in 2003-04. The rates for Australia and Canada were typically higher than those in the United Kingdom. However, the Canadian rate fell to a little over 2% in 2004 and the Australia rate fell to below 3%.

We can use the data on indexed bonds to measure expected inflation rates. To illustrate, we know from Figure 11.4 the real interest rate, r_t, on U.S. indexed bonds of 10-year maturity. We can use data on U.S. Treasury nominal bonds to measure the nominal interest rate, i_t, for the same maturity, 10 years. The difference between i_t and r_t reflects financial market expectations of inflation over the 10 years, that is, it reflects the expected inflation rate, π_t^e, over 10 years. To see this, remember that the expected real interest rate on nominal bonds is given from equation (11.6) by

$$r_t^e \ (\textit{on nominal bonds}) = i_t (\textit{on nominal bonds}) - \pi_t^e .$$

If the expected real interest rate on nominal bonds, r_t^e, equals the (guaranteed) real interest rate, r_t, on indexed bonds,[8] we can substitute r_t on the indexed bonds for r_t^e on the nominal bonds to get

---

[8] This condition would hold if households regarded expected real returns on nominal bonds to be the same as guaranteed real returns on indexed bonds. More generally, uncertainty about the actual real interest

$$r_t\,(on\ indexed\ bonds) = i_t\,(on\ nominal\ bonds) - \pi_t^e\,.$$

We can then rearrange terms to get the expected inflation rate, π_t^e, on the left-hand side:

$$\pi_t^e = i_t\ (on\ nominal\ bonds) - r_t\ (on\ indexed\ bonds). \qquad (11.7)$$

Thus, we can calculate π_t^e by using market data on the two interest rates on the right-hand side.

Figure 11.6 shows as the red graph the 10-year expected inflation rate, π_t^e, computed for the United States from equation (11.7). The green graph shows π_t^e over 30 years, and the blue graph shows π_t^e over 5 years. In 2004, the values for π_t^e were between 2% and 3%. These measures for π_t^e match up reasonably well with the 6-month-ahead expected inflation rates from the Livingston survey, shown by the red graph in Figure 11.2. The values in Figure 11.6 are more variable—and probably more accurate—than the Livingston numbers. However, the Livingston survey has the advantage of providing information back to the late 1940s.

Figure 11.7 shows the analogous measures of π_t^e computed from the 10-year bond data for the United Kingdom, Australia, and Canada. The values for π_t^e were around 8% for the United Kingdom in the early 1980s but decreased substantially thereafter to reach 2-3% in 2003-04. Similar patterns apply to the shorter samples available for Australia and Canada.

---

rate, r_t, on nominal bonds means that the expected real interest rate, r_t^e, on nominal bonds would deviate from the real interest rate, r_t, on indexed bonds.

E. Interest rates on money

We have discussed the nominal and real interest rates on bonds, i_t and r_t. The same analysis applies to money (currency), once we specify that the nominal interest rate on money is zero, rather than i_t. The real interest rate on any asset is the nominal interest rate less the inflation rate, π_t. Hence, for bonds, we had the relation

$$r_t = i_t - \pi_t. \tag{11.5}$$

The analogous condition for money is

$$\begin{aligned} \textit{real interest rate on money} &= \textit{nominal interest rate on money} - \pi_t \\ &= 0 - \pi_t \\ &= -\pi_t. \end{aligned}$$

Thus, if π_t is greater than zero, the real interest rate on money is less than zero. This negative real interest rate signifies that the real value of money erodes over time because of increases in the price level.

As with bonds, we can distinguish the expected real interest rate on money from the actual rate. The expected real interest rate on money is the negative of the expected inflation rate, $-\pi_t^e$.

Remember that the money in our model takes the form of paper currency, which pays a zero nominal interest rate. Most forms of checkable deposits, which are included in the M1 definition of money, pay a positive nominal interest rate. For these types of money, we would calculate the real interest rate just as we do for bonds in equation (11.5). Therefore, the real interest rate on checkable deposits equals the nominal interest rate, i_t, less the inflation rate, π_t.

III. Inflation in the Market-Clearing Model

In chapters 6-8, we considered markets for goods, labor, capital services, and bonds. In chapter 10, we added the condition that the nominal quantity of money equals the nominal quantity demanded. We now extend this framework to allow for inflation. In making this extension, we have two major objectives. First, we want to see how inflation affects our previous conclusions about the determination of real variables. These variables include real GDP, consumption and investment, quantities of labor and capital services, the real wage rate, and the real rental price. The real interest rate is another real variable that can be added to this list. Second, we want to understand how inflation is determined. As part of this analysis, we have to consider the determination of the nominal interest rate.

At this point, we consider fully anticipated inflation, that is, situations in which the inflation rate, π_t, equals the expected rate, π_t^e. We postpone our consideration of unexpected inflation until chapter 15. When the actual inflation rate equals the expected inflation rate, we also have that the real interest rate on nominal bonds, r_t, equals the expected real interest rate, r_t^e.[9]

Our discussion of the cross-country data suggested that inflation would be closely related to money growth. Therefore, we have to extend our framework to allow for growth of money, which we identify with paper currency. The easiest way to allow for this growth is to assume that the government prints new currency and gives it to people. A famous imaginary story about this money creation comes from Milton Friedman (1969,

[9] For indexed bonds, $r_t = r_t^e$ always holds. For nominal bonds, equation (11.7) implies that $r_t = r_t^e$ if $\pi_t = \pi_t^e$.

pp. 4-5). In this story, our public officials stuff a helicopter full of paper currency and fly around dropping money randomly over the countryside. When people pick up the money, they effectively receive a **transfer payment** from the government. Although this story is unrealistic, it provides a simple device for introducing new money into the economy. The important assumption in the story is that each person's transfer is independent of his or her income, money holdings, and so on. Economists refer to these kinds of transfers as **lump-sum transfers**, that is, the amount that a person receives is independent of his or her choices of how much to consume and save, how much to work, how much money to hold, and so on. Therefore, we do not have to consider how people adjust their behavior in order to raise their chances of receiving helicopter drops of cash. We shall find later that more realistic ways of allowing for money creation generate similar results about inflation.

A. Intertemporal-substitution effects

We mentioned that intertemporal-substitution effects for consumption and labor supply involve the expected real interest rate, r_t^e. We are now assuming that actual and expected inflation rates are equal, $\pi_t = \pi_t^e$, so that the actual real interest rate, r_t, equals the expected rate, r_t^e. Therefore, we can use r_t to gauge intertemporal-substitution effects.

B. Bonds and capital

Households still have two forms of earning assets, bonds and ownership of capital. We argued in chapter 6 that the rates of return on these two forms of assets had to be equal; otherwise, households would be unwilling to hold both types. Therefore, when the inflation rate, π, was zero, we got the condition

rate of return on bonds = rate of return from owning capital

$$i = (R/P)\cdot\kappa - \delta(\kappa). \quad (11.8)$$

The expression on the right-hand side, from equation (8.6), gives the rate of return from owning capital. This expression depends on the real rental price, R/P, the utilization rate for capital, κ, and the depreciation rate, $\delta(\kappa)$. Now we have to modify this result to allow for nonzero inflation.

The expression on the right-hand side of equation (11.8) still gives the real rate of return from owning capital. That is, $(R/P)\cdot\kappa$ is the real rental payment received per unit of capital, and $\delta(\kappa)$ is the rate at which capital physically disappears due to depreciation. However, we have to modify the left-hand side of equation (11.8) to replace the nominal interest rate on bonds, i, by the real rate, r. Thus, the revised condition is

real rate of return on bonds = real rate of return from owning capital

$$r = (R/P)\cdot\kappa - \delta(\kappa). \quad (11.9)$$

Another way to get the same answer is to think in terms of nominal rates of return, rather than real rates. Going back to equation (11.8), the left-hand side, i, is the nominal interest rate on bonds. Instead of modifying this term to calculate the real interest rate on bonds, r, we can modify the right-hand side, $(R/P)\cdot\kappa - \delta(\kappa)$, to calculate the nominal rate of return from owning capital. This nominal return has a new component because the nominal value of a unit of capital is the price level, P. When the inflation rate, π, is greater than zero, the nominal value of each piece of capital rises over time at the rate π. Therefore, we have to add π to the real rate of return from owning capital to calculate the nominal rate of return from owning capital:

$$\textit{nominal rate of return from owning capital} = (R/P)\cdot\kappa - \delta(\kappa) + \pi.$$

The condition for equating the return on bonds to that on capital can therefore be written in terms of nominal rates of return as

$$\textit{nominal rate of return on bonds} = \textit{nominal rate of return from owning capital}$$

$$i = (R/P)\cdot\kappa - \delta(\kappa) + \pi. \tag{11.10}$$

This condition is the same as equation (11.9). To see this, move π from the right-hand side to the left-hand side to get

$$i - \pi = (R/P)\cdot\kappa - \delta(\kappa).$$

The left-hand side is the real interest rate on bonds, $r = i - \pi$. Therefore, we again have the result shown in equation (11.9).

C. Interest rates and the demand for money

Recall that the demand for money involved a tradeoff between transaction costs and asset income. By incurring more transaction costs, households could hold less real money balances, M/P, and thereby more real earning assets, $(B/P + K)$. The nominal interest rate on earning assets was i, whereas that on money was zero. Therefore, the level of the nominal interest rate, i, equaled the difference between the nominal interest rate on earning assets (i) and that on money (zero). For this reason, i determined how much interest income was lost by holding money rather than earning assets. An increase in i made the potential loss of interest income more significant and, therefore, motivated households to incur more transaction costs in order to reduce M/P.

This analysis still applies when the inflation rate, π, is different from zero. In this case, the real interest rate on earning assets is $r = i - \pi$, and the real interest rate on money is $-\pi$. Thus, the difference between the two real interest rates is

$$\begin{aligned} &\textit{real interest rate on earning assets} - \textit{real interest rate on money} \\ &= (i - \pi) - (-\pi) \\ &= i. \end{aligned}$$

Thus, the nominal interest rate, i, equals the difference between the two real interest rates. Note that the difference between the real interest rates is the same as the difference between the nominal interest rates, which also equals i. The reason is that we subtract π from each nominal rate to get the real rate. When we take the difference, as in the equation above, the two adjustments for π cancel out.

We have found when we allow for inflation that it is still the nominal interest rate, i, that determines the interest income lost by holding money rather than earning assets. Therefore, as before, an increase in i lowers the real demand for money. We can still describe real money demand by the function used in chapter 10:

$$M^d/P = \pounds(Y, i). \tag{10.2}$$

Notice an important point. It is the real interest rate, r, that has intertemporal-substitution effects on consumption and labor supply. However, it is the nominal interest, i, that influences the real demand for money, M^d/P.

D. Inflation and the real economy

We found in chapter 10 that a change in the nominal quantity of money, M, was neutral. A doubling of M led to a doubling of the price level, P, and to no changes in real variables. The real variables included real GDP, Y, consumption, C, gross investment, I, the real wage rate, w/P, and the real rental price, R/P.

If we allow the nominal quantity of money, M, to grow over time, we shall find that the price level, P, also grows. That is, money growth will create inflation—the inflation rate, π, will be greater than zero. However, we shall find that, at least when the inflation is anticipated, the consequences for real variables tend to be minor.

To see how this works, go back to the setting from chapter 7 where the quantities of capital, K, and labor, L, are fixed, and where we do not allow for a variable utilization rate for capital, κ. The real GDP, Y, is determined by the production function,

$$Y = A \cdot F(\kappa K, L).$$

With fixed κK and L and a given technology level, A, Y is fixed. Hence, Y cannot depend on the inflation rate, π.

The fixed values of κK and L determine the marginal products of capital services and labor, MPK and MPL. As before, the market-clearing real rental price, R/P, equals the MPK, and the market-clearing real wage rate, w/P, equals the MPL. The values for R/P and κ determine the real rate of return from owning capital, $(R/P)\cdot\kappa - \delta(\kappa)$. We know from before that the real interest, r, equals this rate of return:

$$r = (R/P)\cdot\kappa - \delta(\kappa). \tag{11.9}$$

The important point is that we have determined r without considering the inflation rate, π. That is, r is independent of π.

We know that real GDP, Y, is independent of the inflation rate, π. The final issue is the division of Y between consumption, C, and gross investment, I. We want to know whether π affects this division. To answer this question, we have to consider how households determine consumption. Recall from our analysis in chapter 6 that there are two possible ways to affect consumption: the intertemporal-substitution effect and the income effect.

Recall that the intertemporal-substitution effect depends on the real interest rate, r. We know that r is determined from equation (11.6), which does not involve the inflation rate, π. Since π does not affect r, we conclude that π cannot affect consumption through an intertemporal-substitution effect.

The second possibility is that the inflation rate, π, affects consumption, C, through an income effect. For example, if an increase in π makes households feel poorer, they would reduce C. Then, since real GDP, Y, is fixed, gross investment, I, would have to rise. Conversely, if a rise in π makes households feel richer, C would rise, and I would fall. A full analysis of income effects from inflation is complicated. The main point is that rising price make people feel poorer from the standpoint of buying goods but make people feel richer from the standpoint of selling goods. This effect cancels out in the aggregate. For this reason, income effects are typically small enough to neglect. We therefore ignore income effects for our analysis of inflation.

In order to affect consumption, C, the inflation rate, π, has to exert an intertemporal-substitution effect or an income effect. However, we have found that changes in π have neither an intertemporal-substitution effect nor an income effect (as an approximation). Therefore, we conclude that differences in π do not affect C. In this case, differences in π also do not affect gross investment, I.

Our conclusion is that, at least as an approximation, differences in the inflation rate, π, do not matter for a list of real variables: real GDP, Y, consumption, C, gross investment, I, the real rental price, R/P, the real interest rate, r, and the real wage rate, w/P. We reached this conclusion for a case in which the quantities of labor, L, and capital services, κK, were fixed. However, the result still holds when households choose labor supply, L^s, and the utilization rate for capital, κ. When we use the analysis from chapter 8, we find that π does not affect the equilibrium values of L and κ.[10] In that case, the list of unchanged real variables expands to include L and κ.

E. Money growth, inflation, and the nominal interest rate

We want now to examine the linkage between money growth and inflation. We carry out this analysis for given values of the real interest rate, r, and real GDP, Y. By holding these variables fixed, we are making two types of assumptions. First, we are using the results from the previous section, where we found that the inflation rate, π, did not affect r and Y, at least as an approximation. Second, we are assuming that r and Y are not changing over time for other reasons. For example, we could have that Y was growing over time due to increases in the capital stock, K, labor, L, and the technology level, A. These changes, which we analyzed with the Solow model in chapters 3-5, can readily be added to the analysis. The important assumption—the first one—is that variations in the inflation rate, π, do not influence r and Y.

Let ΔM_t be the change in the nominal quantity of money from year t to year $t+1$:

$$\Delta M_t = M_{t+1} - M_t.$$

[10] In Figure 8.3, a change in π would not shift the curve for labor supply, L^s, and therefore would not affect the market-clearing values of the real wage rate, w/P, and the quantity of labor, L. In Figure 8.8, a change in π would not shift the curve for the supply of capital services, $(\kappa K)^s$, and therefore would not affect the market-clearing values of the real rental price, R/P, and the quantity of capital services, κK.

The growth rate of money from year t to year $t+1$, denoted by μ_t, is the ratio of this change in money to the quantity of money:

$$\mu_t = \Delta M_t/M_t. \tag{11.11}$$

For example, if $M_t = 100$ and $\Delta M_t = 5$, the growth rate of money from year t to year $t+1$ is

$$\mu_t = 5/100$$
$$= 0.05 \text{ per year (or 5\% per year)}.$$

We can illustrate how money growth affects inflation by considering a constant growth rate of money. That is, μ_t equals a constant, μ. In this case, equation (11.11) becomes

$$\mu = \Delta M_t/M_t$$
$$= (M_{t+1} - M_t)/M_t.$$

If we multiply through by M_t, we get

$$\mu M_t = M_{t+1} - M_t.$$

If we move M_t from the right-hand side to the left-hand side, combine the terms involving M_t, and then switch the right- and left-hand sides, we get

$$M_{t+1} = (1+\mu)\bullet M_t. \tag{11.12}$$

Thus, the nominal quantity of money rises from year t to year $t+1$ by the factor $1+\mu$. For example, if $M_t = 100$ and $\mu = 0.05$, $M_{t+1} = 1.05\bullet 100 = 105$.

The inflation rate, π_t, for year t is defined as

$$\pi_t = \Delta P_t/P_t$$
$$= (P_{t+1} - P_t)/P_t.$$

If we multiply through by P_t, we get

$$\pi_t P_t = P_{t+1} - P_t.$$

If we move P_t from the right-hand side to the left-hand side, combine the terms involving P_t, and then switch the right- and left-hand sides, we get

(11.13) $$P_{t+1} = (1+\pi_t)\bullet P_t.$$

Thus, the price level rises from year t to year $t+1$ by the factor $1+\pi_t$. For example, if $P_t = 100$ and $\pi_t = 0.05$, $P_{t+1} = 1.05\bullet 100 = 105$.

In chapter 10, we determined the price level from the condition that the real quantity of money, M/P, equaled the real quantity demanded, $£(Y, i)$, where Y is real GDP and i is the nominal interest rate. We found that a once-and-for-all increase in the nominal quantity of money, M, raised the price level, P, in the same proportion. It is therefore reasonable to consider the possibility that, when M_t is growing steadily at the rate μ, the price level, P_t, will also grow at the rate μ. That is, our conjecture is that the inflation rate, π_t, will be constant at the value $\pi = \mu$. In this case, equation (11.13) becomes

(11.14) $$P_{t+1} = (1+\mu)\bullet P_t.$$

We can study how real money balances change over time by dividing equation (11.12) by equation (11.14) to get

$$\frac{M_{t+1}}{P_{t+1}} = \frac{(1+\mu)\bullet M_t}{(1+\mu)\bullet P_t} = \frac{M_t}{P_t}.$$

Hence, when the nominal quantity of money, M_t, and the price level, P_t, grow at the same rate, μ, the level of real money balances, M_t/P_t, does not change over time. Real money balances will therefore equal a constant.

To finish the analysis, we have to ensure that the constant level of real money balances, M/P, equals the real quantity demanded, given by the function $£(Y, i)$. For this equality to hold, we first have to be sure that the real quantity of money demanded is constant. The first determinant of real money demand is real GDP, Y, which we have already assumed to be constant. To assess the second determinant, i, recall that the relation among the real interest rate, the nominal interest rate, and the inflation rate is

(11.5) $$r_t = i_t - \pi_t.$$

Therefore, if we rearrange the terms and replace r_t by the constant r and π_t by the constant μ, we have

(11.15) $$i = r + \mu.$$

Thus, i is constant over time and equal to the sum of the real interest rate, r, and the growth rate of money, μ (which equals the inflation rate, π).

Since the two determinants of money demand, Y and i, are constant, the real quantity of money demanded, given by $£(Y, i)$, is constant. We showed before that the real quantity of money, M_t/P_t, is also constant. Therefore, the only remaining thing to verify is that the two constants are the same. If they are, the real quantity of money will equal the real quantity demanded at every point in time.

Consider, for example, the current year, year 1. We want to ensure that the real quantity of money, M_1/P_1, equals the real quantity demanded, $£(Y, i)$:

(11.16) $$M_1/P_1 = £(Y, i).$$

We can rearrange the terms in this equation to solve out for the price level, P_1:

(11.17) $$P_1 = M_1/£(Y, i).$$

We know everything on the right-hand side because the nominal quantity of money, M_1, is given, and the determinants of real money demand, Y and i, are known. In particular, equation (11.15) implies $i = r + \mu$. Thus, equation (11.17) determines the price level, P_1.

After year 1, the nominal quantity of money, M_t, and the price level, P_t, grow at the same rate, μ. Therefore, real money balances, M_t/P_t, do not change. Hence, for any year t, M_t/P_t equals M_1/P_1, which appears on the left-hand side of equation (11.16). The right-hand side of this equation, $£(Y, i)$, also stays constant over time. We conclude that real money balances, M_t/P_t, equal the real quantity demanded, $£(Y, i)$, in every year t. Thus, we have verified our guess that the inflation rate, π_t, is the constant $\pi = \mu$. As long as year 1's price level, P_1, satisfies equation (11.17), the inflation rate $\pi = \mu$ is consistent with our basic equilibrium condition: the real quantity of money equals the real quantity demanded in every year.

To summarize, the results are as follows:

- The inflation rate equals the constant growth rate of money, $\pi = \mu$.
- Real money balances, M_t/P_t, are constant.
- The nominal interest rate, i, is constant and equal to $r + \pi$, where r is the constant real interest rate, determined as in chapter 8.
- The real demand for money, $£(Y, i)$, is constant, where Y is the constant real GDP, determined as in chapter 8.
- Year 1's price level, P_1, is determined to equate the quantity of real money balances, M_1/P_1, to the real quantity demanded, $£(Y, i)$.

F. Trends in the real demand for money

Thus far, we assumed that the real demand for money, $£(Y, i)$, was fixed. To be more realistic, we should allow for variations in this real demand. Our basic approach would be the same regardless of the source of these variations. For example, sustained growth in real GDP, Y, as analyzed in chapter 5, would lead to steady growth in the real quantity of money demanded. Alternatively, we could allow for continuing improvements in financial technology. If these improvements made currency less useful, the real demand for money might be steadily decreasing.

To take a specific case, suppose that the financial technology and the nominal interest rate, i, are fixed. Assume that real GDP, Y, is rising steadily at the rate γ. The empirical evidence discussed in chapter 10 suggested that the real demand for currency, which is our definition of money, is proportional to Y. In this case, growth in Y at the rate γ leads to growth in the real quantity of money demanded, $\pounds(Y, i)$, at the same rate. Thus, we now assume that this real quantity demanded is rising over time at the constant rate γ.

Now we calculate the rate at which real money balances, M_t/P_t, are growing. We know that, in equilibrium, the growth rate of M_t/P_t has to equal the growth rate of the real quantity demanded, which we know equals γ.

The nominal quantity of money, M_t, is still growing at the constant rate μ, so that

(11.12) $$M_{t+1} = (1+\mu)\bullet M_t.$$

The inflation rate, π, will turn out to be constant but no longer equal to μ. In any event, since the price level, P_t, is growing at the constant rate π, we have from equation (11.13):

(11.18) $$P_{t+1} = (1+\pi)\bullet P_t.$$

If we divide equation (11.12) by equation (11.18), we get

(11.19) $$\frac{M_{t+1}}{P_{t+1}} = \frac{(1+\mu)}{(1+\pi)} \bullet \frac{M_t}{P_t}.$$

Hence, real money balances, M_t/P_t, rise over time by the factor $(1+\mu)/(1+\pi)$.

The growth rate of real money balances is

$$\textit{growth rate of } M_t/P_t = \frac{(M_{t+1}/P_{t+1}) - (M_t/P_t)}{M_t/P_t}.$$

If we substitute for M_{t+1}/P_{t+1} from equation (11.19), we get

$$\textit{growth rate of } M_t/P_t = \frac{\frac{(1+\mu)}{(1+\pi)} \bullet \frac{M_t}{P_t} - \frac{M_t}{P_t}}{\frac{M_t}{P_t}}.$$

Therefore, if we cancel out the M_t/P_t terms, we have

$$\textit{growth rate of } M_t/P_t = (1+\mu)/(1+\pi) - 1.$$

We can simplify the expression on the right-hand side by putting everything over $1+\pi$:

$$\textit{growth rate of } M_t/P_t = \frac{1+\mu-(1+\pi)}{1+\pi}.$$

Then, if we cancel out the 1 and -1 in the numerator, we have

$$\textit{growth rate of } M_t/P_t = (\mu - \pi)/(1+\pi).$$

If π is small relative to 1, the term $1+\pi$ is close to 1, and the last expression is approximately equal to $\mu - \pi$. In fact, if we had carried out our analysis in continuous time, rather than years, the result would have been exactly $\mu - \pi$. Therefore, we use the more accurate result:

$$\textit{growth rate of } M_t/P_t = \mu - \pi. \tag{11.20}$$

That is, the growth rate of real money balances equals the growth rate of nominal money, μ, less the growth rate of prices, π. Thus, if μ is greater than π, M_t/P_t will be rising over time.

The equilibrium condition is still that the quantity of real money balances, M_t/P_t, equals the real quantity demanded, $£(Y, i)$, in every year t. Since $£(Y, i)$ is growing at the rate γ, we must have that the growth rate of M_t/P_t is also γ. Therefore, if we substitute this result into equation (11.20), we get

$$\gamma = \mu - \pi.$$

If we rearrange the terms to solve out for the inflation rate, π, we have

$$\pi = \mu - \gamma. \quad (11.21)$$

Thus, the inflation rate, π, is less than the money growth rate, μ, if the growth rate of real GDP, γ, is greater than zero.[11] However, we again have that an increase in μ by 1% per year leads, for given γ, to an increase in π by 1% per year. That is, consistent with Figure 11.1, variations in money growth rates are still the source of variations in inflation rates.

The new result from equation (11.21) is that variations in the growth rate of real GDP, γ, affect the relationship between the growth rates of money, μ, and price, π. Specifically, a higher γ raises μ - π and, therefore, increases the growth rate of real money balances, M_t/P_t. Figure 11.8 uses the information for 82 countries from Table 11.1 to check out this proposition. The horizontal axis plots γ, from column 4 of the table. The vertical axis has the growth rate of real currency, from column 3 of the table. The graph shows that a higher γ matches up with a higher growth rate of real currency. The correlation between these two variables is 0.72. The slope of the relation is close to one—that is, if γ is higher by 1% per year, the growth rate of real currency is higher by about 1% per year. Therefore, if γ is higher by 1% per year, the inflation rate, π, falls by 1% per year for a given money growth rate, μ.

[11] More generally, we would calculate π by subtracting from μ the growth rate of the quantity of real money demanded, *£(Y, i)*. In our example, the growth rate of *£(Y, i)* equals γ, the growth rate of real GDP.

G. A shift in the money growth rate

We can better understand the relation between the money growth rate and the inflation rate by studying how the economy reacts to a shift in the money growth rate. To simplify, return to the setting in which real GDP, Y, is constant and the real quantity of money demanded, $£(Y, i)$, is not growing over time. Suppose that the nominal quantity of money, M_t, has been growing for a long time at the constant rate μ. Thus, M_t is given by the red line on the left-hand side of Figure 11.9. Since the graph uses a proportionate (logarithmic) scale, the slope of the line equals μ.

We assume that everyone initially expects the monetary authority to keep the quantity of money, M_t, growing forever at the rate μ. In this case, our previous analysis applies, and the inflation rate is the constant

$$\pi = \mu.$$

We show the price level, P_t, as the blue line on the left-hand side of Figure 11.9. The slope of this line equals the inflation rate, $\pi = \mu$. Therefore, the blue line parallels the red line.

Suppose that the real interest rate is the constant r. The nominal interest rate, i, is then given by

$$i = r + \pi.$$

Therefore, if we substitute μ for π, we have

$$i = r + \mu. \tag{11.22}$$

Assume that the monetary authority raises the money growth rate from μ *to* μ' in year T. The right-hand side of Figure 11.9 shows that the quantity of money, M_t, is given after year T by the brown line. This line has the slope μ', which is larger than the slope μ of the red line.

We think of the change in money growth in year T as a surprise—something that people did not anticipate beforehand. However, once the change occurs, we assume that everyone expects the new growth rate of money, μ', to persist forever. Therefore, after year T, the economy is in the same type of situation as before. The only difference is that the money growth rate is μ', rather than μ. Therefore, the inflation rate after year T will equal the money growth rate:

$$\pi' = \mu'.$$

The right-hand side of Figure 11.9 shows that the price level, P_t, follows the green line. The slope of this line equals the inflation rate, $\pi' = \mu'$. Therefore, this slope is the same as that of the brown line.

We know that the inflation rate after year T equals π', which is greater than the inflation rate, π, before year T. Therefore, the slope of the green line in Figure 11.9 is greater than that of the blue line. Notice, however, an important complication in the graph of P_t. The blue line does not intersect the green line in year T. Instead, the figure shows an upward jump in the price level in year T. To understand why this jump occurs we have to think about the determinants of the demand for money.

We know from before that a change in the money growth rate will not affect the real interest rate. Therefore, after year T, the real interest rate will still be r. The nominal interest rate after year T will be

$$i' = r + \pi'.$$

Therefore, if we substitute μ' for π', we have

$$(11.23) \qquad i' = r + \mu' .$$

If we subtract equation (11.22) from equation (11.23) we get

$$(11.24) \qquad i' - i = \mu' - \mu.$$

Thus, the increase in the nominal interest rate, $i' - i$, equals the increase in the money growth rate, $\mu' - \mu$.

The important thing about a rise in the nominal interest rate is that it lowers the real quantity of money demanded, *£(Y, i)*. Therefore, this real quantity demanded is lower after year *T* than before year *T*. Our basic equilibrium condition is that the real quantity of money, M_t/P_t, equals the real quantity demanded, *£(Y, i)*, in every year. Therefore, the real quantity of money, M_t/P_t, must be lower after year *T* than before year *T*.

If we consider the left-hand side of Figure 11.9, we see that, before year *T*, the nominal quantity of money, M_t, grows at the same rate, μ, as the price level, P_t. Therefore, the real quantity of money, M_t/P_t, is constant during this period. Similarly, if we look at the right-hand side of the figure, we see that, after year *T*, M_t, grows at the same rate, μ', as P_t. Hence, M_t/P_t is also constant during this period. We therefore have three facts:

- M_t/P_t is constant before year *T*.
- M_t/P_t is constant after year *T*.
- M_t/P_t after year *T* is lower than that before year *T* (because of the rise in the nominal interest rate from *i* to *i'*).

The way to reconcile these facts is that the price level, P_t, has to jump upward during year T, as shown in Figure 11.9. This rise in P_t lowers real money balances, M_t/P_t, from the value prevailing before year T to that prevailing after year T.

One way to think about the jump in the price level is that the inflation rate is exceptionally high during year T. Specifically, when the money growth rate rises in year T from μ to μ', the inflation rate, π_t, rises in the short run to a value even higher than μ'. The excess of π_t over μ' means that the price level, P_t, rises at a faster rate than the nominal quantity of money, M_t, so that real money balances, M_t/P_t, decline.

In our case, the interval of exceptionally high inflation is concentrated in a short period, essentially an instant of time. In other words, the transition from relatively low inflation, $\pi = \mu$, to relatively high inflation, $\pi' = \mu'$, takes place overnight. More generally, the essence of the transition is that the level of real money balances, M_t/P_t, has to decrease—because higher inflation leads to a higher nominal interest rate, which reduces the real quantity of money demanded. Thus, the transition always requires π_t to exceed μ_t over some interval. The exact pattern of π_t during this transition and the length of the transition depend on the details of the model. The transition is not always limited to a single year or to an instant of time, as shown in Figure 11.9.

In one modification of the model, households adjust their real demand for money downward only gradually in response to an increase in the nominal interest rate, i_t. This gradual adjustment makes sense if households have to change their underlying cash-management plans or payment frequencies in order to hold lower real money balances. In this case, the transition to a lower level of real money balances, M_t/P_t, is stretched out. As real money demand falls gradually, M_t/P_t also decreases gradually. The period of gradually falling M_t/P_t corresponds to an extended transition in which the inflation rate, π_t, exceeds the growth rate of money, μ_t.

As another example, households may know in advance that the monetary authority is planning to shift from relatively low money growth, μ, to relatively high money growth, μ'. That is, households may expect before year T that money growth and inflation will be higher from year T onward. In this case, some of the unusually high short-run inflation would occur prior to year T. That is, the expectation of higher inflation in the long run leads to higher inflation in the short run, even before the rise in money growth occurs. The higher short-run inflation arises because the anticipated future inflation will decrease the real quantity of money demanded before year T.

Expectations about future changes in money growth and inflation have sometimes been at the center of political campaigns. For example, in the 1890s, William Jennings Bryan campaigned for the U.S. presidency on a program of easy money. In essence, he advocated a higher growth rate of money, μ, based on free coinage of silver as a supplement to the monetary role of gold. Bryan's defeat probably lowered expectations of future money growth and inflation.

As another example, in 1980, Ronald Reagan campaigned for president partly on the promise of getting tough on inflation, which was then high in the United States. Reagan's defeat of Jimmy Carter at the polls likely lowered expectations of future money growth and inflation.

Another, more extreme, case concerns the post-World War I hyperinflation in Germany, which we examine in a box in a later section. The end of the hyperinflation occurred in November 1923. However, people anticipated prior to November that a monetary reform was coming and that this reform would entail lower money growth rates, μ, and inflation rates, π. Studies of the German hyperinflation have shown that this anticipation reduced π before November 1923, even though the reduction in μ had not yet occurred.[12]

H. Government revenue from printing money

We have assumed, thus far, that the monetary authority prints new money (currency) and gives it to people as transfer payments. More realistically, governments use the **revenue from printing money** to pay for a variety of expenditures. Governments do not usually use this revenue to finance Milton Friedman's imaginary helicopter drops of cash!

If we think (realistically) of the monetary authority as part of the government, the government's nominal revenue from printing money between years *t* and *t+1* equals the change in the nominal quantity of money:

$$\begin{aligned} \textit{nominal revenue from printing money} &= M_{t+1} - M_t \\ &= \Delta M_t. \end{aligned}$$

To calculate the real value of this revenue, we have to divide by P_{t+1}, the price level for year $t+1$:

$$\textit{real revenue from printing money} = \Delta M_t/P_{t+1}.$$

Recall that the growth rate of money, μ_t, is defined as

$$\mu_t = \Delta M_t/M_t. \qquad (11.11)$$

Therefore, if we multiply through by M_t, we get $\mu_t \cdot M_t = \Delta M_t$. If we substitute $\mu_t \cdot M_t$ for ΔM_t in the formula for real revenue, we get

[12] See Flood and Garber (1980) and Lahaye (1985).

$$real\ revenue\ from\ printing\ money = \mu_t \cdot (M_t/P_{t+1}).$$

The term on the far right-hand side, M_t/P_{t+1}, is approximately the level of real money balances, M_t/P_t. (If we had done the analysis in continuous time, rather than years, the term would be exactly the level of real money balances.) Therefore, we can approximate the real revenue from printing money as

$$(11.25) \qquad real\ revenue\ from\ printing\ money \approx \mu_t \cdot (M_t/P_t)$$
$$= money\ growth\ rate\ times\ level\ of\ real\ money\ balances.$$

We know that a higher money growth rate, μ_t, leads to a higher inflation rate, π_t, and a higher nominal interest rate, i_t. We also know that the higher i_t reduces the real quantity of money demanded and, therefore, lowers the level of real money balances, M_t/P_t. Thus, an increase in the money growth rate, μ_t, has two opposing effects on the real revenue from printing money in equation (11.25). The rise in μ_t raises the real revenue, but the decrease in M_t/P_t reduces the real revenue. The net effect depends on how much real money demand falls in response to a rise in i_t.

As an example, suppose that real money balances, M_t/P_t, were initially equal to 100 and that μ_t then doubled from 5% to 10%. The real revenue from printing money would rise on net unless M_t/P_t fell to below 50. That is, real revenue would increase unless M_t/P_t fell by more than 50%. More generally, the real revenue rises unless the decrease in real money demand is proportionately larger than the increase in the money growth rate. This condition holds empirically except for the most extreme cases. For example, during the German hyperinflation, which we consider in the next section, the condition was apparently violated only when the money growth rate, μ_t, approached the astronomical value of 100% per month between July and August 1923. Until then, the government extracted more real revenue by printing money at a faster rate.

In normal times for most countries, the government obtains only a small portion of its revenue from printing money. In 2003, for example, the Federal Reserve obtained $22 billion from this source. This amount constituted 1.2% of total federal receipts and 0.2% of GDP. These figures are typical for the main developed countries.

In a few high-inflation countries, the revenue from money creation became much more important. For example, in Argentina from 1960 to 1975, money creation accounted for nearly half of government revenue and to about 6% of GDP. Some other countries in which the revenue from printing money was important were Chile (5% of GDP from 1960 to 1977), Libya (3% of GDP from 1960 to 1977), and Brazil (3% of GDP from 1960 to 1978).

During the German hyperinflation and other hyperinflations, money creation became the primary source of government receipts.[13] The amounts obtained in some hyperinflations approached 10% of GDP, which appears to be about the maximum obtainable from printing money. In the German case, from 1920 to 1923, there was a close month-to-month connection between the volume of real government spending and the money growth rate. Thus, the variations in the money growth rates—and, hence, the inflation rates—were driven by shifts in real government spending.[14] Much of the government spending at the time went to reparations payments to the victors in World

[13] Commenting on this situation, Keynes (1923) said "A Government can live for a long time, even the German Government or the Russian Government, by printing paper money. That is to say, it can by this means secure the command over real resources—resources just as real as those obtained by taxation."
[14] See Hercowitz (1981) for a detailed analysis.

War I. Therefore, the reductions in these payments after November 1923 were a major factor in the ending of the German hyperinflation.

Money and prices during the German hyperinflation

The post-World War I German hyperinflation provides a great laboratory experiment for studying the interplay between money growth and inflation.[15] From 1920 to 1923, the inflation rates ranged from near zero to over 500% per month! (When talking about hyperinflations, economists measure inflation rates per month, rather than per year.) The data suggest that relatively small changes occurred in aggregate real variables, such as real GDP. Therefore, the quantity of real money demanded was not much affected by changes in real GDP.

When inflation rates are volatile, as in post-World War I Germany, it is impossible to predict accurately the real interest rate on loans that prescribe nominal interest rates. Therefore, this type of lending tends to disappear. For this reason, we have no useful measures of the nominal interest rate during the German hyperinflation. In this environment, the best measure of the cost of holding money is the expected inflation rate, π_t^e. This rate determined how much income people lost by holding money rather than consuming or holding a

[15] Not surprisingly, this episode has fascinated many economists. The classic study was by Bresciani-Turroni (1937). Cagan (1956) studied the German hyperinflation along with six others: Austria, Hungary, Poland, and Russia after World War I and Greece and Hungary after World War II. The Hungarian experience after World War II seems to be the all-time record hyperinflation. In this case, the price level rose by a factor of 3×10^{25} over the 13 months from July 1945 to August 1946. For a discussion, see Bomberger and Makinen (1983).

durable good that maintained its real value over time. Empirical studies have estimated π_t^e by assuming that it adjusted gradually to changes in the actual inflation rate, π_t.

Table 11.3 summarizes the behavior of the money growth rate, μ_t (based on currency in circulation), the inflation rate, π_t, and real money balances, M_t/P_t, in Germany from 1920 to 1925. In most cases, the table shows μ_t and π_t over six-month intervals. The level of M_t/P_t pertains to the ends of these intervals.

At the beginning of 1920, the money growth rate, μ_t, and the inflation rate, π_t, were already at the very high rate of 6% per month. Then μ_t declined at the beginning of 1921 to an average of less than 1% per month. Notice that, as our model predicts, π_t fell by more than μ_t, so that real money balances, M_t/P_t, rose by about 20% from early 1920 to early 1921. (The level of M_t/P_t in early 1920 was about the same as that before the war in 1913.)

From late 1921 through the end of 1922, the money growth rate, μ_t, rose to about 30% per month. During this period, the inflation rate, π_t, exceeded μ_t. Hence, by the end of 1922, real money balances, M_t/P_t, fell to about a quarter of the level from early 1920.

From the end of 1922 through mid 1923, the money growth rate, μ_t, was extremely high but was no longer trending upward—it averaged around 40% per month. In this period, the inflation rate, π_t, was also around 40% per month,

Table 11.3			
Money Growth and Inflation during the German Hyperinflation			
Period	μ_t	π_t	M_t/P_t (end of period)
2/20-6/20	5.7	6.0	1.01
6/20-12/20	3.0	1.1	1.13
12/20-6/21	0.8	0.1	1.18
6/21-12/21	5.5	8.4	0.99
12/21-6/22	6.5	12.8	0.68
6/22-12/22	29.4	46.7	0.24
12/22-6/23	40.0	40.0	0.24
6/23-10/23	233	286	0.03
Reform period			
12/23-6/24	5.9	-0.6	0.44
6/24-12/24	5.3	1.4	0.56
12/24-6/25	2.0	1.6	0.57
6/25-12/25	1.2	0.4	0.60

Source: *Sonderhefte zür Wirtschaft und Statistik*, Berlin, 1925.

Note: The nominal quantity of money, M_t, is an estimate of the total circulation of currency. Until late 1923, the figures refer to total legal tender, most of which constituted notes issued by the Reichsbank. Later, the data include issues of the Rentenbank, private bank notes, and various "emergency moneys." However, especially in late 1923, many unofficial emergency currencies, as well as circulating foreign currencies, are not counted. The numbers are normalized so that M_t in 1913 is set to 1.0. The price level, P_t, is an index of the cost of living, based on 1913 = 1.0. Column 1 shows the period for the data. Column 2, in percent per month, is the growth rate of money, μ_t, over the period shown. Column 3, also in percent per month, is the inflation rate, π_t, over the period shown. Column 4 is the level of real money balances, M_t/P_t, at the end of each period. The value of M_t/P_t for 1913 is normalized to equal 1.0. Therefore, the values shown for M_t/P_t are relative to the level in 1913.

so that real money balances, M_t/P_t, remained at around 25% of the early-1920 level. However, in late 1923, the hyperinflation built to its climax, with μ_t reaching 300-600% per month in October and November. Real money balances, M_t/P_t, reached their low point in October at about 3% of the level from early 1920.

A major monetary reform occurred in Germany in November 1923. This reform included the introduction of a new type of currency, a promise not to print new money beyond a specified limit to finance government expenditures, changes in government spending and taxes, and a commitment to back the value of the new currency by gold.[16] This reform led to a sharp cutback in the money growth rate, μ_t, and the inflation rate, π_t, after December 1923. During 1924, μ_t averaged around 5% per month. However, π_t was much lower, less than 1% per month. The excess of μ_t over π_t allowed for the rebuilding of real money balances, M_t/P_t, which increased from 3% of the early-1920 level in October 1923 to 56% of that level by December 1924. A substantial part of the rise in M_t/P_t occurred during the reform months of November and December 1923, when the government introduced large amounts of its new currency.

In 1925, the money growth rate, μ_t, fell below 2% per month, and the inflation rate, π_t, remained at about 1% per month. Thus, real money balances,

[16] For discussions of the reform, see Bresciani-Turroni (1937), Sargent (1982), and Garber (1982). Sargent's analysis deals also with the ends of the hyperinflations in Austria, Hungary, and Poland in the early 1920s. He stresses the rapidity with which inflations can be ended once governments make credible commitments to limit money creation.

M_t/P_t, rose slowly to reach 60% of the early-1920 level by the end of 1925. Although π_t stayed low for the remainder of the 1920s, M_t/P_t did not reattain its early-1920 level. Perhaps this gap reflected a long-lasting negative influence of the hyperinflation on the real demand for money.

IV. Summing up

We argued that sustained inflation requires sustained money growth, and we found this pattern in the cross-country data. An increase in the money growth rate by 1% per year is associated with an increase in the inflation rate by 1% per year. An increase in the growth rate of real GDP by 1% per year is associated with an increase by 1% per year in the growth rate of real money balances. In other words, a higher growth rate of real GDP lowers the inflation rate for a given growth rate of money.

The nominal interest rate exceeds the real interest rate by the rate of inflation. Conventional nominal bonds specify the nominal interest rate, whereas indexed bonds specify the real interest rate. Using U.S. and other data on conventional and indexed bonds, we can measure expected real interest rates and expected inflation rates.

Intertemporal-substitution effects depend on the real interest rate, whereas the demand for money depends on the nominal interest rate. We showed in the model how an increase in the money growth rate leads one-for-one to a higher inflation rate and a higher nominal interest rate. As an approximation, a change in money growth does not affect a list of real variables, including real GDP, consumption, investment, and the real interest rate. However, higher money growth leads to lower real money balances and, typically, to greater real revenue for the government. The real revenue from money creation is small in most circumstances but is high in hyperinflations.

Questions and Problems

Mainly for review

11.1. Which of the following statements is correct?

a. A constant rate of increase in the price level, *P*, will lead to a continuous rise in the nominal interest rate, *i*.

b. A continuous increase in the inflation rate, π, will lead to a continuous rise in the nominal interest rate, *i*.

11.2. Define the real interest rate, *r*. Why does it differ from the nominal interest rate, *i*, in the presence of inflation?

11.3. Why does the actual real interest rate, *r*, generally differ from the expected real interest rate, r_t^e? How does this relation depend on whether bonds prescribe the nominal or real interest rate?

11.4. What is the Livingston survey of inflationary expectations? What are the pluses and minuses of using this type of information to measure the expected rate of inflation, π_t^e?

11.5. Why is it the real interest rate, rather than the nominal rate, that has intertemporal-substitution effects on consumption and saving? Does the same argument apply to intertemporal substitution of labor supply?

Problems for discussion

11.x. Money growth and inflation

Suppose that the money-demand function takes the form

$$M^d/P = £(Y, i) = Y \cdot \Psi(i).$$

That is, for a given nominal interest rate, i, a doubling of real GDP, Y, doubles the real demand for money, M^d/P.

a. Consider the relation across countries between the growth rate of money (currency), μ, and the inflation rate, π, as shown in Figure 11.1. How does the growth rate of real GDP, $\Delta Y/Y$, affect the relationship between μ and π?

b. What is the relation between μ and π for a country in which the nominal interest rate, i, has increased?

c. Suppose that the expected real interest rate, r_t^e, is given. What is the relation between μ and π for a country in which the expected inflation rate, π_t^e, has increased?

11.x. Statistical relations between money growth and inflation

Students who have studied econometrics and have access to a statistical package can do the following exercise.

a. Use the data in Table 11.1 to run a regression of the inflation rate, π, on a constant and the growth rate of money (currency), μ. What is the estimated coefficient on μ and how should we interpret it? What is the meaning of the constant term?

b. Run a regression of the growth rate of real money balances, $\mu - \pi$, on the growth rate of real GDP, $\Delta Y/Y$ and a constant. What is the estimated coefficient on $\Delta Y/Y$ and how should we interpret it?

c. Suppose that we add the variable $\Delta Y/Y$ to the regression run in part a. What is the estimated coefficient on $\Delta Y/Y$ and how should we interpret it?

11.x. Effects on the nominal interest rate

What would be the effect on the price level, P, and the nominal interest rate, i, from the following events?

a. A once-and-for-all increase in the nominal quantity of money, M.

b. A once-and-for-all increase in the growth rate of money, μ.

c. A credible announcement that the growth rate of money, μ, will rise beginning one year in the future.

11.x. Seasonal variations in money

Suppose that the real quantity of money demanded is relatively high in the fourth quarter of each year and relatively low in the first quarter. Assume that there is no seasonal pattern in real interest rates.

a. Suppose that there were no seasonal in the nominal quantity of money, M. What would the seasonal be for the price level, P, the inflation rate, π, and the nominal interest rate, i?

b. What seasonal behavior for the nominal quantity of money, M, would eliminate the seasonal patterns in P, π, and i?

11.x. Interest-rate targeting

Suppose that the monetary authority wants to keep the nominal interest rate, i, constant. Assume that the real interest rate, r, is fixed. However, the real demand for money, M^d/P, shifts around a great deal.

a. How should the monetary authority vary the nominal quantity of money, M, if the real demand for money, M^d/P, increases temporarily? What if the real demand increases permanently?

b. How does the price level, P, behave in your answers to part a.? What should the monetary authority do if it wants to dampen fluctuations of P, as well as maintaining a constant nominal interest rate, i?

11.x. Money growth and government revenue

Can the government always increase its real revenue from printing money by raising the money growth rate, μ? How does the answer depend on the responsiveness of real money demand, M^d/P, to the nominal interest rate, i?

11.x. Prepayment of mortgages and callability of bonds

Mortgages typically allow the borrower to make early payments ("pre-payments") of principal. Sometimes the mortgage contract specifies a prepayment penalty, and sometimes there is no penalty. Similarly, long-term bonds (though typically not those issued by the U.S. government) sometimes allow the issuer to prepay the principal after a prescribed date and with a specified penalty. When the bond issuer exercises this option to prepay, he or she is said to "call" the bond. Bonds that allow this prepayment are said to be "callable" or to have a "call provision."

a. When would a borrower want to prepay (or call) his or her mortgage or bond? Would we see more prepayments when the nominal interest rate, i, had unexpectedly increased or decreased?

b. From the late 1970s until 1982, banks and savings & loan associations were eager for customers to prepay their mortgages. Why was this the case? Later on, customers wanted to prepay. Why did they want to do so?

c. Suppose that the year-to-year fluctuations in nominal interest rates become larger. (These fluctuations—or volatility—were particularly great from the mid 1970s through the early 1980s.) From the standpoint of a borrower, how does this change affect the value of having a prepayment option—that is, callability—in his or her mortgage or bond?

11.x. Rational expectations and measures of expected inflation

How would the hypothesis of rational expectations help us to measure the expected inflation rate, π_t^e? What seem to be the pluses and minuses of this approach?

11.x. Indexed bonds

a. Consider first a one-year nominal bond that costs $1000. The bond pays after one year the principal of $1000 plus an interest payment of $50. What is the one-year nominal interest rate on the bond? What are the actual and expected one-year real interest rates on the bond? Why is the nominal interest rate known but the real rate uncertain?

b. Consider now a one-year indexed bond (such as the U.S. Treasury's TIPS—or Treasury Inflation-Protected Securities). Suppose that the bond costs \$1000. One year later, the nominal principal of the bond is adjusted to be $\$1000 \cdot (1 + \pi)$, where π is the actual inflation rate over the year. Then the bond pays off the adjusted principal of $\$1000 \cdot (1 + \pi)$ plus an interest payment of, say, 3% of the adjusted principal. What is the one-year real interest rate on the indexed bond? What are the actual and expected one-year nominal interest rates on the bond? Why is the real rate known but the nominal rate uncertain?

c. Can you think of other ways to design indexed bonds? Are the nominal and real interest rates both uncertain in some cases?

11.x. A case of counterfeiting

In 1925 a group of swindlers induced the Waterlow Company, a British manufacturer of bank notes, to print and deliver to them 3 million pounds' worth of Portuguese currency (escudos). Since the company also printed the legitimate notes for the Bank of Portugal, the counterfeit notes were indistinguishable from the real thing (except that the serial numbers were duplicates of those from a previous series of legitimate notes). Before the fraud was discovered, 1 million pounds' worth of the fraudulent notes had been introduced into circulation in Portugal. After the scheme unraveled (because someone noticed the duplication of serial numbers), the Bank of Portugal made good on the fraudulent notes by exchanging them for newly printed, valid notes. The Bank subsequently sued the Waterlow Company for damages. The company was found liable, but the key question was the amount of damages. The Bank argued that the damages were 1 million pounds (less funds collected from the swindlers). The other side contended that the Bank suffered only negligible real costs in having to issue an additional 1 million pounds' worth of the money to redeem the fraudulent notes. (Note that the currency was purely a paper issue, with no convertibility into gold or anything else.) Thus, the argument was that the only true costs to the Bank were the expenses for paper and printing. Which side do you think was correct? (The House of Lords determined in 1932 that 1 million pounds was the correct measure. For discussions of this fascinating episode in monetary economics, see Roy Hawtrey [1932] and Murray Bloom [1966].)

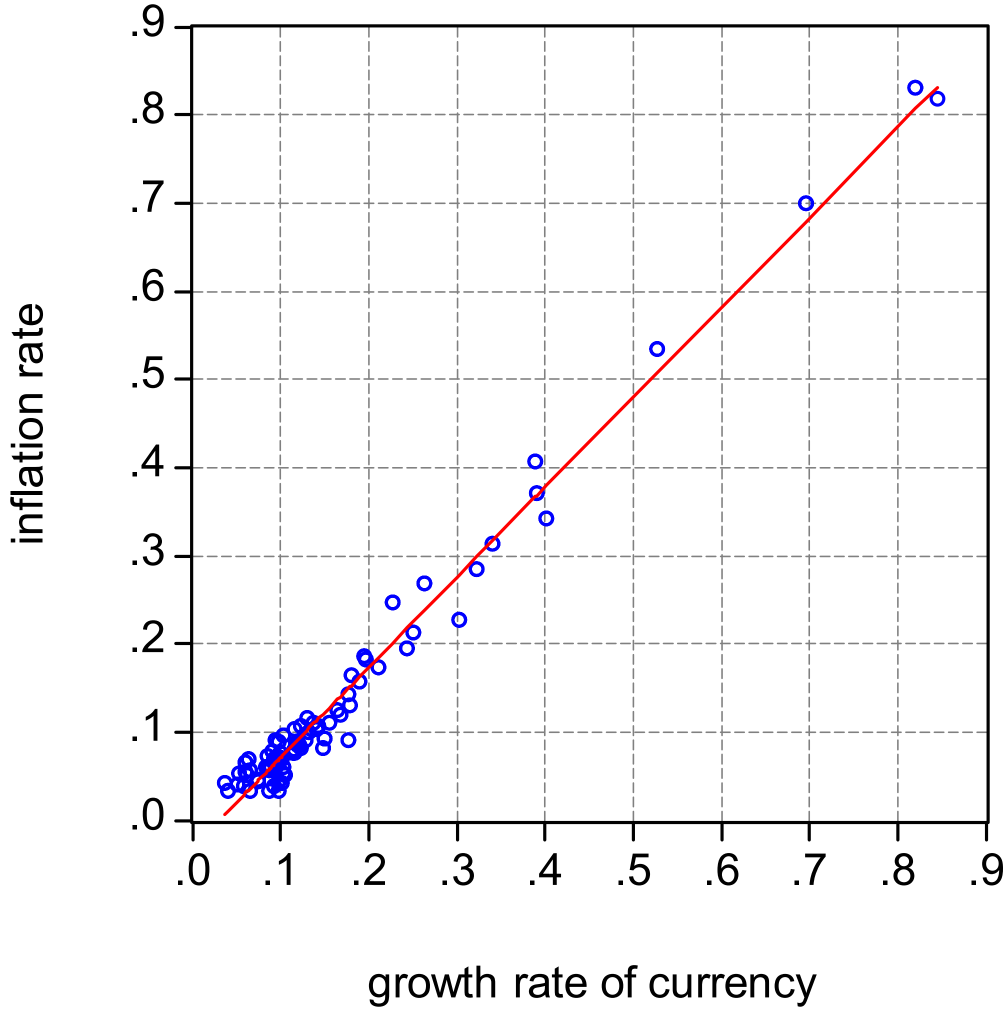


Figure 11.1

Inflation Rate and Growth Rate of Nominal Currency for 82 Countries, 1960-2000

The vertical axis has the inflation rate from 1960 to 2000, based on consumer price indexes. The horizontal axis has the growth rate of nominal currency from 1960 to 2000. The two variables have a remarkably strong positive association—the correlation is 0.99. The slope of the relation is close to one—that is, an increase in the growth rate of nominal currency by 1% per year is associated with an increase in the inflation rate by about 1% per year.

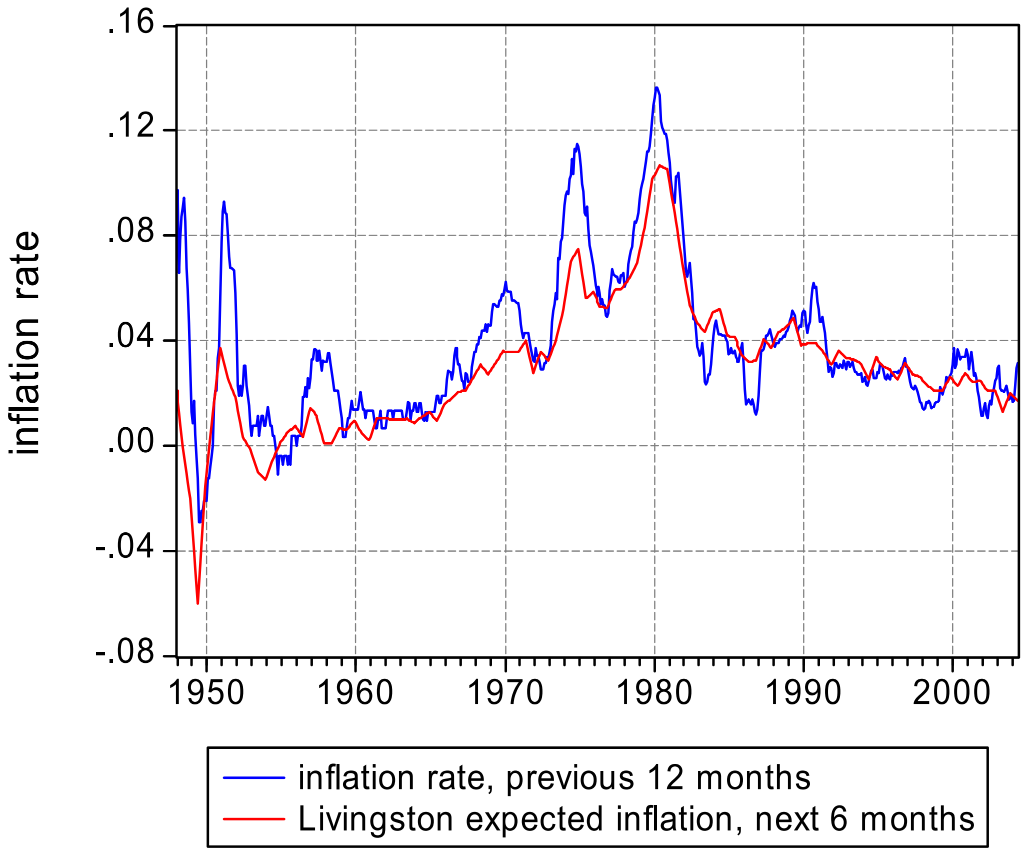


Figure 11.2

Actual and Expected Inflation Rates in the United States

The blue graph shows the inflation rate over the previous 12 months, computed from the consumer price index (CPI). The red graph shows the expected inflation rate. These expectations, formed six to eight months in advance, are from the Livingston survey, available from the Federal Reserve Bank of Philadelphia.

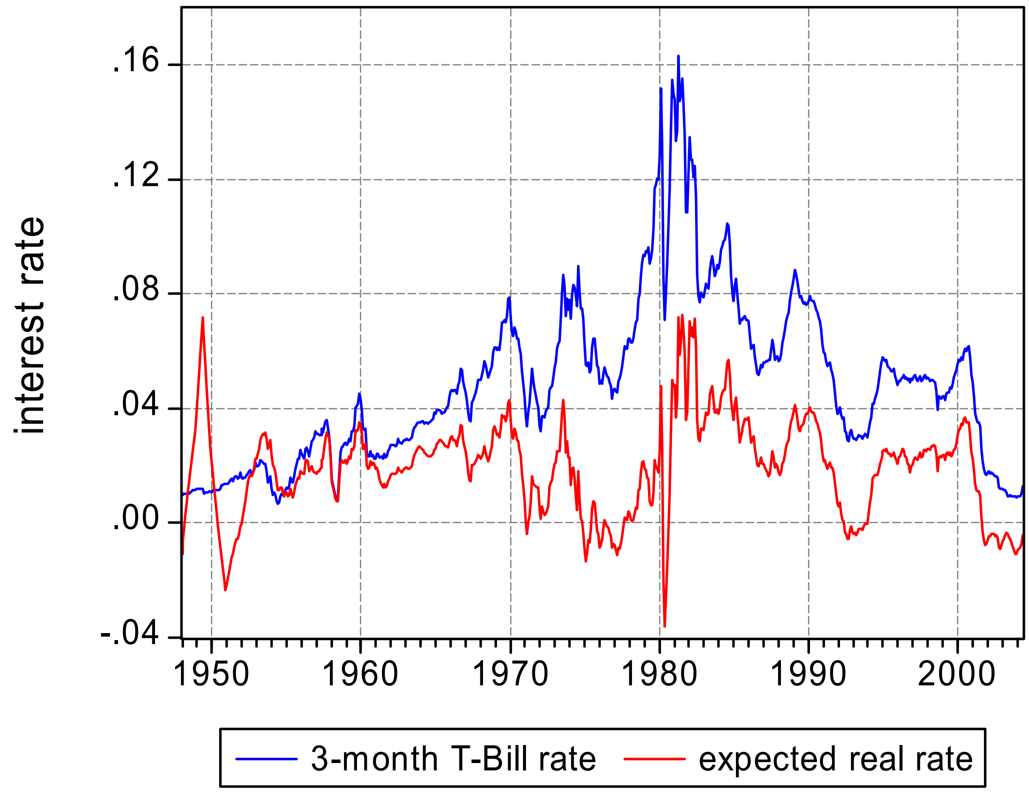


Figure 11.3

Nominal and Expected Real Interest Rates in the United States

The 3-month U.S. Treasury Bill rate, shown as the blue graph, is a nominal interest rate. We compute the expected real interest rate by subtracting an expected inflation rate, given by the Livingston survey measure shown in Figure 11.2. The resulting expected real interest rate is the red graph.

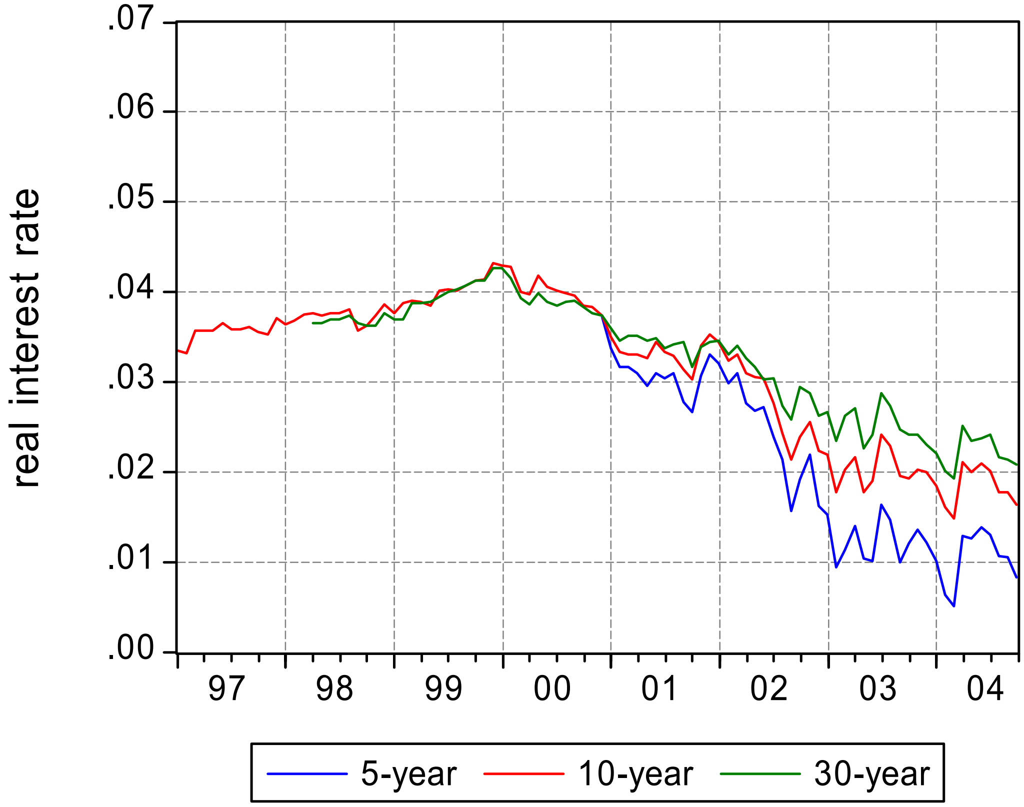


Figure 11.4

Real Interest Rates on U.S. Indexed Bonds

The graphs show the real interest rate on inflation-protected U.S. Treasury bonds (indexed bonds). The blue graph is for 5-year maturity, the red graph for 10-year maturity, and the green graph for 30-year maturity. Data are from Global Financial Data, Inc.

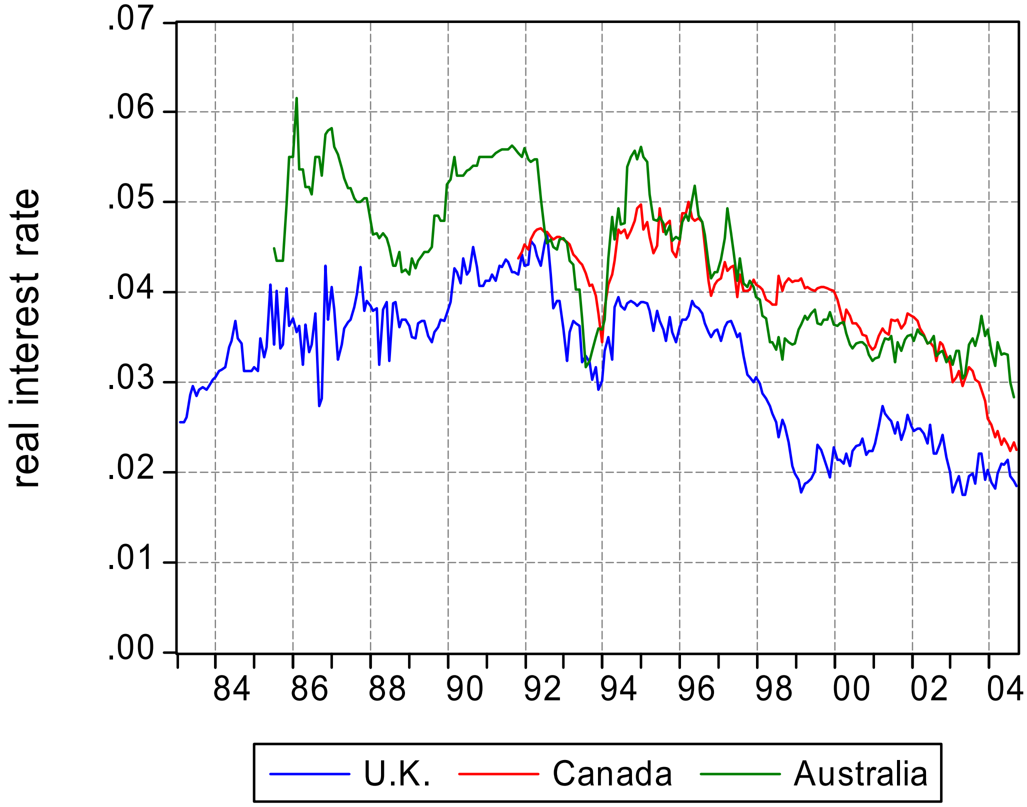


Figure 11.5

Real Interest Rates on Indexed Bonds in the U.K., Canada, and Australia

The graphs show the real interest rate on inflation-protected 10-year government bonds (indexed bonds). The blue graph is for the United Kingdom, the red graph for Canada, and the green graph for Australia. Data are from Global Financial Data, Inc.

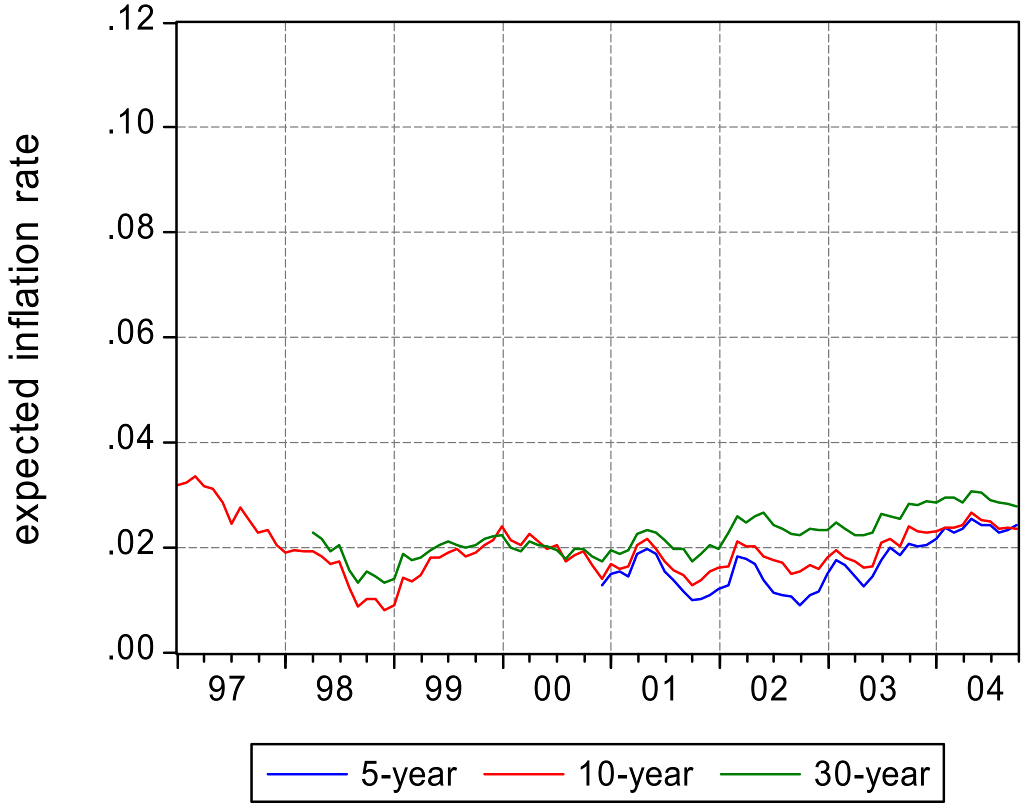


Figure 11.6

U.S. Expected Inflation Rates, Based on Indexed Bond Yields

We compute expected inflation rates by taking the nominal interest rate on nominal U.S. Treasury bonds and subtracting the real interest rate on indexed U.S. Treasury bonds (from Figure 11.4). See equation (11.7). The blue graph is based on 5-year bonds, the red graph on 10-year bonds, and the green graph on 30-year bonds. Thus, the graphs measure expected inflation rates over 5 years, 10 years, and 30 years, respectively.

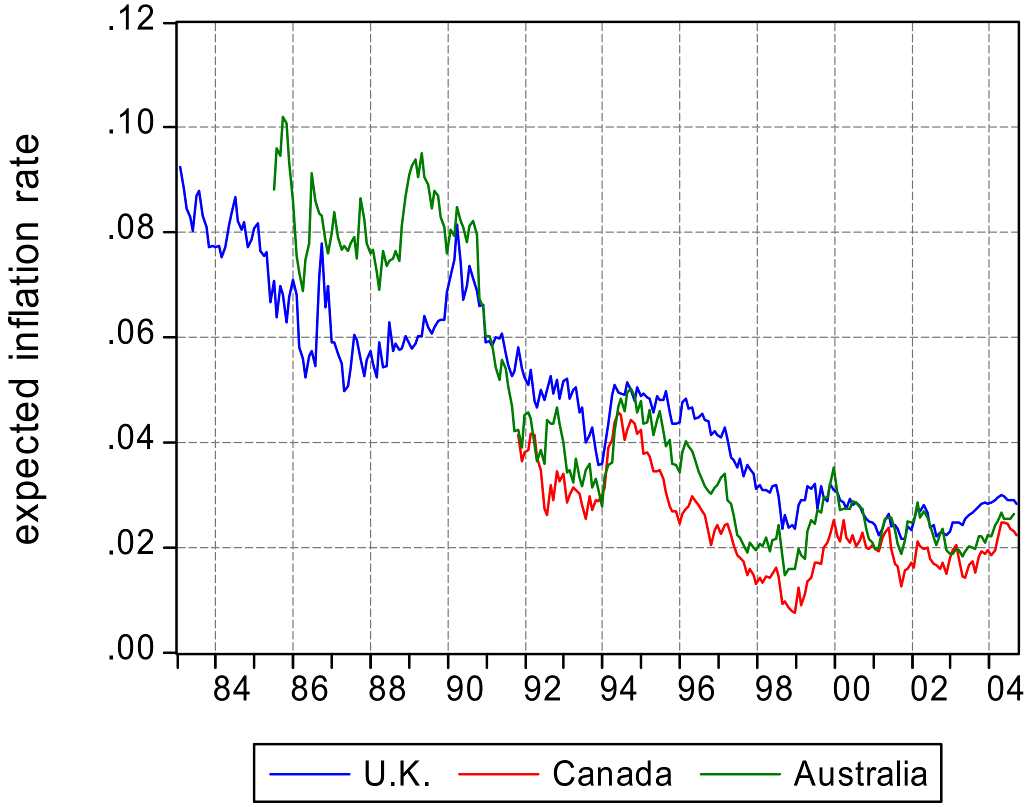


Figure 11.7

Expected Inflation Rates, Based on Indexed Bond Yields, in the U.K., Canada, and Australia

We compute expected inflation rates by taking the nominal interest rate on nominal government bonds and subtracting the real interest rate on indexed government bonds (from Figure 11.5). See equation (11.7). These calculations are based on 10-year bonds. The blue graph is for the United Kingdom, the red graph for Canada, and the green graph for Australia. These graphs measure expected inflation rates over 10 years.

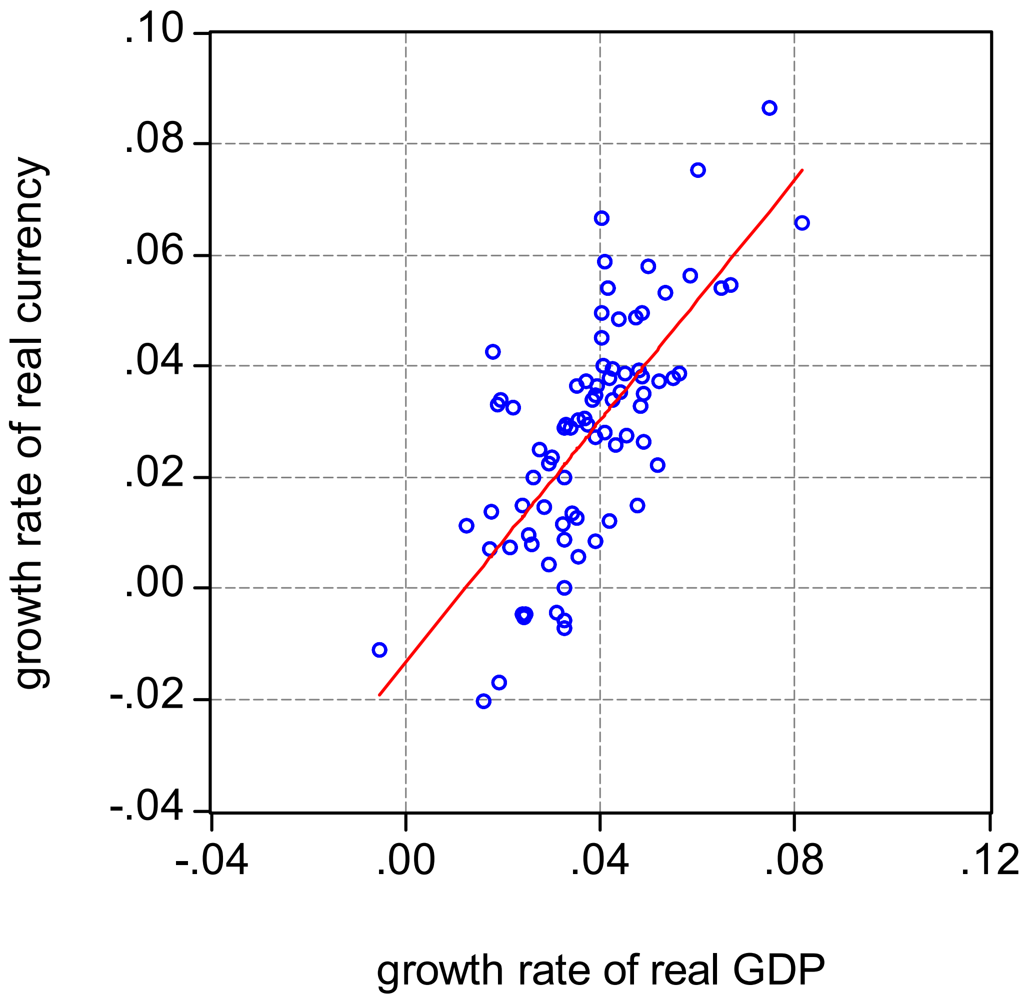


Figure 11.8

Growth Rate of Real Currency and Growth Rate of Real GDP for 82 Countries, 1960-2000

The horizontal axis has the growth rate of real GDP from 1960 to 2000. The vertical axis has the growth rate of real currency from 1960 to 2000. The two variables have a correlation of 0.72. The slope is close to one—that is, an increase in the growth rate of real GDP by 1% per year is associated with an increase in the growth rate of real currency by about 1% per year.

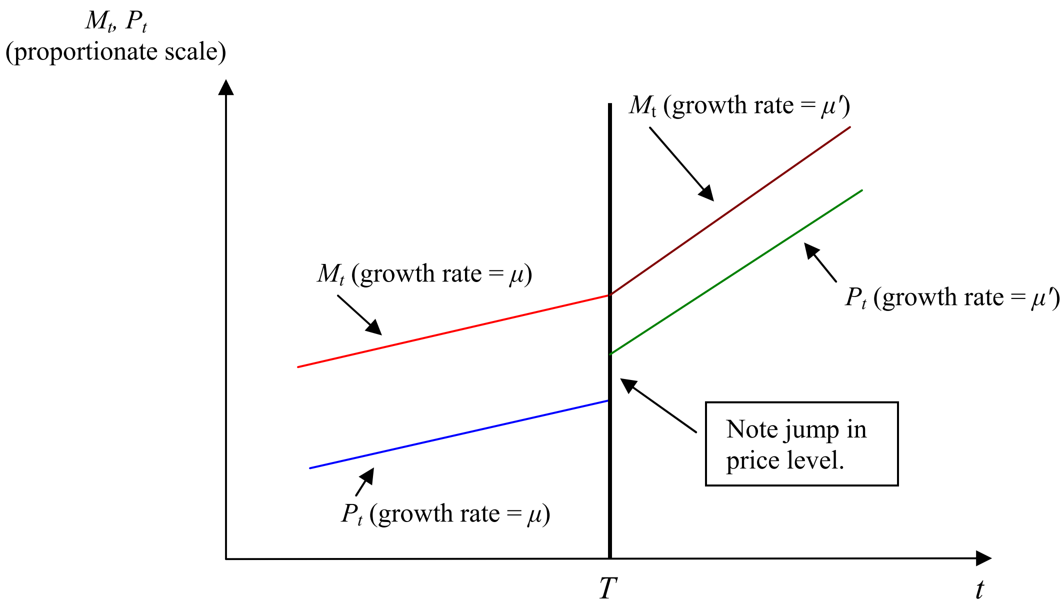


Figure 11.9

Effect of an Increase in the Money Growth Rate on the Path of the Price Level

The red line shows that the nominal quantity of money, M_t, grows at the constant rate μ before year T. After year T, M_t grows along the brown line at the higher rate μ'. The blue line shows that the price level, P_t, grows at the same rate as money, μ, before year T. After year T, P_t grows along the green line at the same rate as money, μ'. The price level, P_t, jumps upward during year T. This jump reduces real money balances, M_t/P_t, from the level prevailing before year T to that prevailing after year T.

January 12, 2005

Chapter 12

Government Expenditure

Up to now, the government had very limited functions in our model. We considered only lump-sum transfers financed by money creation. Now we allow for the government's purchases of goods and services. In the national accounts, these outlays are called government consumption and investment.[1] We assume in this chapter that these expenditures are financed by lump-sum taxes, which are analogous to the lump-sum transfers that we considered before. We also continue to assume that the transfers are lump sum. In the next chapter, we allow for more realistic systems of taxes and transfers. It is useful to start with data on government expenditure for the United States and other countries.

I. Data on Government Expenditure

Figures 12.1 and 12.2 show the evolution of government expenditure in the United States from 1929 to 2004. Excluding the wartime experiences, total nominal government spending, expressed as a ratio to nominal GDP, went from 0.10 in 1929 to 0.19 in 1940, 0.24 in 1950, 0.28 in 1960, 0.32 in 1970, 0.33 in 1980, 0.34 in 1990, 0.31

[1] The difference between government consumption and investment and government purchases is that the former category includes an estimate of depreciation on parts of the government's capital stock. In the model, the government owns no capital. Hence, depreciation of the government's capital stock is zero, and government investment is also zero. Therefore, in the model, government purchases are the same as government consumption.

in 2000, and 0.33 in 2004. Thus, over the last three decades, total government spending has remained close to one-third of GDP.

One part of government outlays is for purchases of goods and services (or government consumption and investment[2]). Figure 12.1 shows that the federal part of these outlays went from 0.02 relative to GDP in 1929 to 0.06 in 1940, 0.09 in 1950, 0.12 in 1960, 0.11 in 1970, 0.09 in 1980, 0.09 in 1990, 0.06 in 2000, and 0.07 in 2004.

Except during the Great Depression from 1933 to 1940, the dominant part of federal consumption and investment was defense spending (shown separately in the figure). Defense outlays were 0.101 relative to GDP in 1960 but fell from there to a low point of 0.057 in 1979, before rising during the defense buildup of the Reagan administration to 0.074 in 1986. Then the ratio fell steadily, perhaps because of the "peace dividend" associated with the end of the cold war, to 0.038 in 2000. Subsequently, the defense ratio turned upward again, reaching 0.047 in 2004. The figure also shows the local peaks in the ratio of defense spending to GDP during wartime: 0.43 in 1944 (World War II), 0.15 in 1952-53 (Korean War), 0.10 in 1967-68 (Vietnam War), and 0.047 in 2004 (wars in Afghanistan and Iraq).

Consumption and investment by state and local governments was 0.07 relative to GDP in 1929 and 1950, then rose to a peak of 0.13 in 1975. About half of this increase reflected growth in outlays for education, which went from 0.024 relative to GDP in 1952 to 0.056 in 1975. After 1975, the ratio of state & local consumption and investment to GDP remained nearly stable, varying between 0.11 and 0.12. The educational part of this spending was also a stable share of GDP in this period.

[2] In the national accounts, government consumption includes an estimate of depreciation of publicly owned capital. This depreciation does not correspond to current government purchases of goods and services.

Figure 12.2 shows the evolution of government transfer payments.[3] The ratio of total transfers to GDP has been trending upward: it went from 0.008 in 1929 to 0.024 in 1940, 0.061 in 1950, 0.055 in 1960, 0.079 in 1970, 0.103 in 1980, 0.106 in 1990, 0.113 in 2000, and 0.123 in 2004.

At the federal level, the largest expansion was in the main social security program, OASDI (old age, survivors, and disability insurance), and Medicare (which pays for medical expenses of the elderly). The ratio of OASDI expenditure to GDP went from zero in 1940 to 0.043 in 1980, then stayed roughly constant through 2003. The ratio of Medicare spending to GDP went from zero in 1965 to 0.025 in 2003.

State and local transfers include welfare payments in the form of family assistance. However, the main increase has been in Medicaid, which pays for health outlays by poor persons. The ratio of these expenditures to GDP went from zero in 1958 to 0.024 in 2003. In 2003, three programs—OASDI, Medicare, and Medicaid—represented 9.1% of the GDP.

Table 12.1 shows ratios of total government expenditure to GDP for 51 countries for 2000-2002. The countries listed are those for which data were available on a broad concept of government spending. This concept includes expenditure by all levels of government. The ratios shown range from 0.13 for Hong Kong to 0.53 for Hungary and Sweden. The median is 0.37, and the United States at 0.32 was somewhat below the median.

[3] The transfers shown in Figure 12.2 include subsidies but exclude payments from the federal government to state & local governments.

II. The Government's Budget Constraint

Now we extend the model to allow for the government's purchases of goods and services. Let G_t represent this spending in real terms for year t. In our previous analysis, we considered two forms of private spending on goods and services: consumption, C_t, and gross investment, I_t. The total of the three terms, $C_t + I_t + G_t$, is the aggregate real spending on goods and services in year t.

The second form of government expenditure, mentioned in chapter 11, is transfer payments. Let V_t represent the government's nominal expenditure on transfers. We can calculate the real value of this spending by dividing by the price level, P_t, to get V_t/P_t.

We assumed in chapter 11 that the government's only revenue came from printing money. The real value of this revenue for year t is $(M_t - M_{t-1})/P_t$, where M_t is the nominal quantity of money in year t.[4] Now we assume that the government also levies taxes on households. These taxes might apply to businesses, but remember that, in the model, the households own and run the businesses. Let T_t be the aggregate nominal taxes collected by the government in year t. The real amount of these taxes is T_t/P_t.

The **government's budget constraint** equates the government's total expenditure to its total revenue. We can write this constraint in real terms for year t as

Key equation (government budget constraint):

real government expenditure = real government revenue

(12.1) $$G_t + V_t/P_t = T_t/P_t + (M_t - M_{t-1})/P_t.$$

[4] In the United States, the revenue from printing money accrues directly to the Federal Reserve. The Fed then turns over most of this revenue to the U.S. Treasury. In the model, we are consolidating the Fed with the government.

Note that we have not yet introduced public debt in the model. Therefore, the government's expenditure on the left-hand side does not include net interest payments, and the government's revenue on the right-hand side does not include the proceeds from issue of public debt. We make the extensions to include interest payments and public-debt issue in chapter 14.

We mentioned in chapter 11 that the real revenue from printing money, $(M_t - M_{t-1})/P_t$, is usually a minor part of overall government receipts. We shall find it convenient to neglect the revenue from printing money, and we can do this by returning to the case in which the nominal quantity of money, M_t, is constant. In this case, we can substitute $M_t – M_{t-1} = 0$ in equation (12.1) to get

real government expenditure = real taxes

$$G_t + V_t/P_t = T_t/P_t. \quad (12.2)$$

In chapter 11, transfer payments took the form of helicopter drops of cash that households picked up. The important feature of these transfers was that they were lump sum—that is, the amount an individual household received did not depend on that household's income, consumption, money holdings, and so on. We continue to assume in this chapter that the real transfers, V_t/P_t, are lump sum. That is, the individual transfers do not depend on an individual's characteristics.

We assume, in this chapter, that the real taxes, T_t/P_t, are also lump sum. That is, the taxes that an individual household pays are assumed to be independent of that household's income, consumption, and so on. This assumption is unrealistic. In the real world, elaborate tax laws specify the relation of a household's taxes to its income, consumption, and so on. There are lots of things that individuals can do—including hiring accountants, working less, underreporting income, and exploiting tax loopholes—to lower their taxes. These possibilities imply important substitution effects from the tax system on labor supply, investment, and even the number of children. Although we want to study these substitution effects, we shall find it convenient to neglect them provisionally in order to isolate the effects from government expenditure. That is why we assume lump-sum taxes in this chapter. In chapter 13, we consider substitution effects from realistic types of taxes. That analysis also allows us to consider substitution effects from real-world transfer programs.

III. Public Production

We assume that the government uses its purchases of goods and services, G_t, to provide services to households and businesses. We assume that the government delivers these services free of charge to the users. In most countries, public services include national defense, enforcement of laws and private contracts, police and fire protection, elementary & secondary schooling and some portions of higher education, parts of health services, highways, parks, and so on. The range of governmental activities has expanded over time, although this range varies from one country or U.S. state to another.

We could model public services as the output from the government's production function. The inputs to this function would be the government-owned stock of capital, labor services from public employees, and materials that the government buys from the private sector. To simplify, we neglect government production and assume instead that the government buys final goods on the goods market. That is, the government's purchases, G_t, add to the demand for goods by private consumers, C_t, and investors, I_t.

In effect, we are assuming that the government sub-contracts all of its production to the private sector. In this setup, public investment, publicly owned capital, and government employment are always zero. Ultimately, we would get different answers by allowing for public production only if the government's production function—that is, its technology and management capability—differed from that of the private sector. Otherwise, it would not matter whether the government buys final goods, as we assume, or, instead, buys capital and labor inputs to produce things itself.

IV. Public Services

We have to take a position on the uses of the services that the government provides. One possibility is that these services yield utility for households. Examples are parks, libraries, school lunch programs, subsidized health care and transportation, and the entertaining parts of the space program. These public services may substitute closely for private consumption. For example, if the government buys a student's lunch at school, the student does not have to buy his or her own lunch.

Another possibility is that public services are inputs to private production. Examples include the provision and enforcement of laws and contracts, aspects of national defense, government-sponsored research & development programs, fire and police services, and regulatory activities. In some cases, public services are close substitutes for private inputs of labor and capital. For example, the government's police services may substitute for guards hired by a private company. In other cases—including infrastructure activities such as the provision of a legal system, national defense, and perhaps transportation facilities—the public services are likely to raise the marginal products of private inputs.

We shall find it convenient to begin with the case in which public services have zero effect on utility and production. Interpreted literally, we are assuming that the government buys goods on the goods market and then throws them into the ocean. We consider later how the conclusions are modified by allowing government services to be useful.

V. Households' Budget Constraints

The government's taxes and transfers will affect households' budget constraints. We can start with the budget constraint worked out in chapter 8. That analysis neglected inflation, that is, the price level, P_t, was constant over time. We simplify by returning to this case. Note that our assumption of a constant P_t is consistent with our assumption of a constant nominal quantity of money, M_t. Neither of these unrealistic assumptions has an important effect on our analysis of government expenditure.

From chapter 8, we have that the household budget constraint at a point in time is

$$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L^s + i\cdot(B/P + K). \tag{8.1}$$

Thus, the sum of consumption, C, and real saving, $(1/P)\cdot\Delta B + \Delta K$, on the left-hand side equals the sum of real wage income, $(w/P)\cdot L^s$, and real asset income, $i\cdot(B/P + K)$, on the right-hand side. Note that the asset income, $i\cdot(B/P + K)$, depends on the nominal interest rate, i. However, since we are assuming that the inflation rate, π, is zero, the real interest rate, r, equals the nominal rate, i. We shall find it useful to replace i by r because then the analysis will still be valid when we allow for π to be nonzero. If we apply equation (8.1) to year t and replace i by r, we get

(12.3) $$C_t + (1/P)\cdot\Delta B_t + \Delta K_t = (w/P)_t\cdot L_t^s + r_{t-1}\cdot(B_{t-1}/P + K_{t-1}),$$

where $\Delta B_t = B_t - B_{t-1}$ and $\Delta K_t = K_t - K_{t-1}$.

The inclusion of the government sector leads to two modifications of the budget constraint in equation (12.3). First, year t's real taxes, T_t/P, appear as a subtraction from real income on the right-hand side. That is, one unit more of real taxes means one unit less of real disposable income. Second, year t's real transfers, V_t/P, enter as an additional form of real income on the right-hand side. Therefore, the households' budget constraint for year t becomes

(12.4) $$C_t + (1/P)\cdot\Delta B_t + \Delta K_t = (w/P)_t\cdot L_t^s + r_{t-1}\cdot(B_{t-1}/P + K_{t-1}) + V_t/P - T_t/P.$$

The new term is the net of real transfers over real taxes, $V_t/P - T_t/P$, on the right-hand side.

We showed in chapters 6 and 8 how to apply the one-year budget constraint to many years. The result, when we modify equation (8.2) to replace the nominal interest rate, i_t, by the real interest rate, r_t, is

(12.5) $$C_1 + C_2/(1+r_1) + C_3/[(1+r_1)\cdot(1+r_2)] + \ldots = (1+r_0)\cdot(B_0/P + K_0)$$
$$+ (w/P)_1\cdot L_1^s + (w/P)_2\cdot L_2^s/(1+r_1) + (w/P)_3\cdot L_3^s/[(1+r_1)\cdot(1+r_2)] + \ldots$$

The left-hand side has the present value of consumption. The right-hand side has the real value of the initial earning assets at the start of year 1 plus the present value of real wage income.

When we extend the one-year budget constraint to include taxes and transfers, as we did in going from equation (12.3) to equation (12.4), we will get the corresponding extension to the multi-year budget constraint, shown in equation (12.5). The right-hand side will add the present value of real transfers net of real taxes. Thus, the extended version of equation (12.5) is

Key equation (household budget constraint with transfers and taxes):

$$(12.6) \quad C_1 + C_2/(1+r_1) + \ldots = (1+r_0)\cdot(B_0/P + K_0) + (w/P)_1\cdot L_1^s + (w/P)_2\cdot L_2^s/(1+r_1) + \ldots$$
$$+ (V_1/P - T_1/P) + (V_2/P - T_2/P)/(1+r_1) + (V_3/P - T_3/P)/[(1+r_1)\cdot(1+r_2)] + \ldots$$

The new term on the right-hand side is the present value of real transfers net of real taxes,

$$(12.7) \quad (V_1/P - T_1/P) + (V_2/P - T_2/P)/(1+r_1) + (V_3/P - T_3/P)/[(1+r_1)\cdot(1+r_2)] + \ldots$$
$$= \textit{present value of real transfers net of real taxes}.$$

A lower present value of real transfers net of real taxes lowers the overall resources available to households. Our analysis from chapter 6 predicts that households would react just as they would to any other loss of income. In particular, the income effects predict reductions in consumption, C_t, and leisure in each year. The decrease in leisure implies an increase in labor supply, L_t^s, in each year.

Our analysis from chapter 6 tells us that the strength of the income effect depends on whether an increase in real transfers net of real taxes is temporary or permanent. For a temporary change, we can consider an increase in year 1's term, $V_1/P - T_1/P$, while holding fixed the terms $V_t/P - T_t/P$ for other years. In this case, the present value of real transfers net of real taxes rises in equation (12.7) but by only a small amount. Therefore, we predict small increases in C_t and small decreases in L_t^s for each year. In contrast, if the rise in $V_t/P - T_t/P$ applies to all years, the present value of real transfers net of real taxes will increase in equation (12.7) by a large amount. Therefore, we predict large increases in C_t and large decreases in L_t^s in each year.

VI. Permanent Changes in Government Purchases

We now use the model to assess the economic effects from a permanent change in government purchases. Recall that Figure 12.2 showed the U.S. data on government purchases, expressed as a ratio to GDP. The present analysis would not apply to large temporary changes, which show up most clearly for defense expenditure during World War II and the Korean War. The analysis does apply to most of the other variations in government purchases that show up in the figure. The reason is that, empirically, most changes in government purchases have been long-lasting, in the sense that predicted future levels of real spending changed by about as much as current real spending.

A. A permanent change in government purchases: theory

Suppose that government purchases, G_t, rise by one unit in each year. Since we are considering the same change for each year, we can simplify the notation by dropping the year subscript, t. In this case, the government's budget constraint from equation (12.2) implies

$$G + V/P = T/P.$$

Therefore, we can rearrange the terms to get a formula for real transfers net of real taxes:

$$V/P - T/P = -G. \tag{12.8}$$

If G rises by one unit in each year, $V/P - T/P$ falls by one unit in each year. Hence, for households, each year's disposable real income falls by one unit. The income effects predict, accordingly, a decrease in each year's consumption, C, and an increase in each year's labor supply, L^s.

We can get the main results by neglecting the changes in labor supply, L^s. That is, we assume that each year's L^s equals a constant, L. We reconsider this assumption in chapter 13, where we allow also for realistic forms of taxes.

Consider the income effect on consumption, C. Since households have one less unit of real disposable income in each year, we predict that the decrease in C in each year will be roughly by one unit. This prediction follows from the result in chapter 8 that the propensity to consume out of a permanent change in income would be close to one.

The real GDP, Y, is given by the production function:

$$Y = A \cdot F(\kappa K, L). \tag{12.9}$$

This formulation allows, as in equation (8.3) of chapter 8, for a variable capital utilization rate, κ, so that κK is the quantity of capital services. We are assuming that the capital stock, K, is fixed in the short run. We are also assuming that the technology level, A, and the quantity of labor input, L, are fixed.

The demand for capital services comes from the schedule for the marginal product, MPK. As in Figure 8.4 of chapter 8, the MPK decreases as the quantity of capital services, κK, rises. Profit-maximizing businesses demand the quantity of capital services $(\kappa K)^d$ that equates the MPK to the real rental price, R/P. Therefore, $(\kappa K)^d$ is a downward-sloping function of R/P. We show this demand curve as the red graph in Figure 12.3.

Since the capital stock, K, is given, the supply of capital services, $(\kappa K)^s$, varies only because of changes in the capital utilization rate, κ. As in Figure 8.8 of Chapter 8, the chosen κ and, hence, the supply of capital services, $(\kappa K)^s$, is an upward-sloping function of the real rental price, *R/P*. We show this supply curve as the blue graph in Figure 12.3.

The important observation is that an increase in government purchases, *G*, does not shift the curves for the demand or supply of capital services. The demand curve does not shift because the rise in *G* does not affect the schedule for the marginal product of capital services, MPK. The supply curve does not shift because, first, *K* is given, and, second, the change in *G* does not affect the choice of capital utilization rate, κ, for a given real rental price, *R/P*. (See section II.B of chapter 8.) Since the demand and supply curves do not shift in Figure 12.3, we conclude that the market-clearing values of the real rental price, *R/P*, and the quantity of capital services, κK, do not change.

Since the quantity of capital services, κK, is unchanged and since we have assumed that the technology level, *A*, and the quantity of labor input, *L*, are fixed, the production function in equation (12.9) implies that real GDP, *Y*, is unchanged. Hence, a permanent increase in government purchases does not affect *Y*.

Consider the real interest rate, *r*. We know that *r* equals the real rate of return on owning capital:

$$r = (R/P)\cdot\kappa - \delta(\kappa). \tag{12.10}$$

The expression on the right-hand side comes from equation (8.6) of chapter 8. The term $(R/P)\cdot\kappa$ is the real rental income per unit of capital (allowing for a variable utilization rate, κ). Since *R/P* and κ are unchanged, the real interest rate, *r*, is the same. That is, a permanent rise in government purchases does not affect *r*.

Now we turn to the labor market. The demand for labor comes from the schedule for the marginal product of labor, MPL. As in Figure 8.3 of chapter 8, the MPL decreases as the quantity of labor, L, rises. Profit-maximizing businesses demand the quantity of labor L^d that equates the MPL to the real wage rate, w/P. Therefore, L^d is a downward-sloping function of w/P. We show this demand curve as the red graph in Figure 12.4.

We discussed in chapter 8 how an increase in the real wage rate, w/P, could induce households to increase their quantity of labor supplied, L^s. However, we are assuming for now that labor supply, L^s, is a constant, L. Therefore, Figure 12.4 shows L^s as a vertical line at L.

The permanent increase in government purchases, G, does not shift the labor-demand curve, L^d, in Figure 12.4. The reason is that the rise in G does not affect the schedule for the marginal product of labor, MPL. To get this answer, we need the result that the quantity of capital services, κK, is unchanged. If capital services had changed, the schedule for the MPL would be different.

We are assuming that the increase in government purchases does not affect labor supply, L^s, which is fixed at L. Given this assumption, nothing shifts in Figure 12.4. Hence, the market-clearing real wage rate, $(w/P)^*$, does not change, and the market-clearing quantity of labor remains at $L^* = L$. We conclude that a permanent increase in government purchases does not affect the real wage rate, w/P.

To complete the analysis, we have to go back to household choices of consumption, C. We found from the income effect that a permanent rise in government purchases, G, would reduce C in each year by roughly one unit. To find the full effect on current consumption, we also have to consider whether any substitution effects are operating. The intertemporal-substitution effect depends on the real interest rate, r.

However, since r does not change, the intertemporal-substitution effect does not operate. Another possible substitution effect involves consumption and leisure, but we have assumed that the quantity of labor and, hence, the quantity of leisure, is fixed. In any event, this substitution effect depends on the real wage rate, w/P, which does not change.

Our conclusion is that, to determine the change in current consumption, C, we have to consider only the income effect. As already noted, if government purchases rise permanently by one unit, the income effect causes C to decline by roughly one unit. Therefore, our prediction is that a permanent increase in government purchases causes a decrease by about the same magnitude in C. To put it another way, the rise in G **crowds out** C roughly one-to-one.

To find the response of gross investment, I, recall that real GDP, Y, equals the sum of consumption, C, gross investment, I, and government purchases, G:

$$Y = C + I + G. \tag{12.11}$$

In the present case, Y is unchanged, G rises by one unit, and C falls by about one unit. Therefore, equation (12.11) tells us that the changes in C and G offset each other and, thereby, allow I to remain unchanged. We conclude that a permanent increase in G does not affect gross investment, I.

To sum up, we found that a permanent increase in government purchases, G, reduces consumption, C, roughly one-to-one. The variables that do not change include real GDP, Y, gross investment, I, the quantity of capital services, κK, the real rental price, R/P, the real interest rate, r, and the real wage rate, w/P.

Useful public services

We did not yet consider that the government may use its purchases, G, to provide useful public services. We study here the case in which these services provide utility for households. For example, the government might provide free or subsidized school lunches or transportation or concerts in the park. We assume that these publicly provided services combine with private consumer expenditure to determine overall household utility. For example, utility depends on transportation, one part of which is provided by the government.

To be concrete, suppose that each unit of government purchases, G, is equivalent in terms of utility to λ units of private consumption, C. We assume that $\lambda \geq 0$ applies. The case $\lambda = 1$ is particularly interesting because it means that each unit of G is equivalent in terms of utility to one unit of C.[5]

Recall from equation (12.4) that the household budget constraint, when written without year subscripts, is

$$C + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L^s + r\cdot(B/P + K) + V/P - T/P.$$

We can add λG to each side of the equation to get

$$(C + \lambda G) + (1/P)\cdot\Delta B + \Delta K = (w/P)\cdot L^s + r\cdot(B/P + K) + V/P - T/P + \lambda G. \quad (12.12)$$

[5] The case $\lambda < 1$ means that each unit of resources that goes through the government provides less utility than each unit of private consumer spending. This case might apply because the lack of market incentives makes government operations relatively inefficient. We could instead have $\lambda > 1$ if there are scale benefits in the provision of public goods.

This specification is useful because $C + \lambda G$ on the left-hand side is the effective flow of consumer services. That is, $C + \lambda G$ combines private consumption, C, with the utility received from public services, λG. The new term on the right-hand side, λG, is the value of the free or subsidized public services. That is, λG is implicitly a form of household income. When we add λG on the right-hand side, we get households' full income, which includes the implicit value of public services.

We assumed before that government purchases, G, rose by one unit in each year. We found from the government's budget constraint in equation (12.8) that the difference between real transfers and real taxes, $V/P - T/P$, fell by one unit in each year. Since this budget condition still applies, we get that the combination of the last three terms on the right-hand side of equation (12.12) changes by

$$\Delta(V/P - T/P + \lambda G) = \Delta(V/P - T/P) + \Delta(\lambda G)$$

$$= -1 + \lambda,$$

where we used the conditions $\Delta(V/P - T/P) = -1$ and $\Delta(\lambda G) = \lambda$. If $\lambda < 1$, full income declines by $1 - \lambda$ units when G rises by one unit. If, instead, $\lambda > 1$, full income rises by $\lambda - 1$ units when G rises by one unit.

Since full income changes by $-1 + \lambda$ units in each year, we would expect households' full consumption, $C + \lambda G$, also to change by about $-1 + \lambda$ units in each year. That is, for a permanent change in full income, we predict that the propensity to change full consumption will be close to unity. To get the change in *C*, we can use the condition

$$\Delta(C + \lambda G) = -1 + \lambda.$$

If we separate out the two changes on the left-hand side, we get

$$\Delta C + \lambda \cdot \Delta G = -1 + \lambda.$$

Then, if we substitute $\Delta G = 1$ and cancel out λ on each side, we get

$$\Delta C = -1.$$

This answer is the same one that we got when we assumed that public services were useless ($\lambda = 0$). An increase in *G* by one unit still crowds out *C* by about one unit. The only difference from before is that, the higher λ, the happier households will be when the government expands *G*.

B. The cyclical behavior of government purchases

One prediction from the model is that long-lasting changes in real government purchases would not have much impact on real GDP. We mentioned before that most changes in real government purchases in the United States fit the assumption of being long-lasting. The main contrary examples come from military expenditure during wartime. These outlays are mostly temporary, tending to last as long as wars.[6] In the

[6] However, in some cases, such as World War II and the Korean War, the effect on defense expenditure was partly long-lasting, tending to persist even after the wars were over.

U.S. data, if we look after the end of the Korean War in 1953, the variations in war-related expenditure were relatively minor. Therefore, the model predicts that the fluctuations in real government purchases that occurred after 1953 should bear little relation to the fluctuations in real GDP.

To test this proposition, Figure 12.5 uses our standard approach of comparing the cyclical part of a variable—in this case, government consumption & investment—with the cyclical part of real GDP. The variability of government consumption & investment is similar to that of real GDP. From 1954.1 to 2004.3, the standard deviations of the logs of the cyclical components were 0.017 and 0.016, respectively. However, the two variables were virtually uncorrelated—the correlation was -0.01. This result is consistent with our prediction that variations in real government purchases would not have much impact on real GDP.

VII. Temporary Changes in Government Purchases

Now we analyze temporary changes in real government purchases. We begin with the model, then study applications to wartime experiences.

A. A temporary change in government purchases: theory

Assume now that year 1's real government purchases, G_1, rise by one unit, while those for other years, G_t, do not change. That is, in year 1, everyone expects that G_t in future years will be unchanged. We can think of this case as representing a war that begins at the start of year 1 and is expected to end by the start of year 2. More generally, the expected duration of a war would differ from one year and, hence, the anticipated interval of heightened government purchases might be greater or less than a year.

The government's budget constraint from equation (12.8) implies for year t:

(12.13) $$V_t/P - T_t/P = -G_t.$$

Therefore, in year 1, the net of real transfers over real taxes, $V_1/P – T_1/P$, falls by one unit. Hence, households have one unit less of real disposable income in year 1. In subsequent years, $V_t/P - T_t/P$ and, hence, real disposable income do not change. Thus, the difference from a permanent rise in government purchases is that the expected real disposable income in future years is unchanged. Our analysis from chapter 8 predicts that households would spread their reduced disposable income in year 1 over reduced consumption, C_t, in all years. Therefore, the effect on year 1's consumption, C_1, will be relatively small. In other words, the propensity to consume out of a temporary change in income is greater than zero but much less than one.

Now we go back to dropping the time subscripts, with each variable implicitly applying to the current year, year 1. Much of the analysis of a temporary change in government purchases is the same as that worked out for a permanent change. Suppose, as before, that we neglect any changes in labor supply, so that $L^s = L$. In Figure 12.3, we still have that the change in government purchases does not affect the schedule for the marginal product of capital services, MPK. Therefore, the change does not shift the demand curve for capital services, $(\kappa K)^d$. The change in government purchases also does not affect the way that suppliers of capital services choose their utilization rate, κ. Therefore, for a given stock of capital, K, the supply curve for capital services, $(\kappa K)^s$, does not shift. Since neither curve shifts in Figure 12.3, we conclude, as before, that the real rental price, R/P, the capital utilization rate, κ, and the quantity of capital services, κK, do not change.

Since capital services, κK, do not change and labor is fixed at L, we know from the production function,

$$(12.9) \qquad Y = A \cdot F(\kappa K, L),$$

that real GDP, Y, does not change. Since the real rental price, R/P, and the capital utilization rate, κ, do not change, we also have that the real interest rate, r, stays the same. This result follows from the formula:

$$(12.10) \qquad r = (R/P) \cdot \kappa - \delta(\kappa).$$

For the labor market, we still have that the change in government purchases does not affect the schedule for the marginal product of labor, MPL. Therefore, in Figure 12.4, the change in government purchases does not shift the labor demand curve, L^d. Since labor supply, L^s, is fixed at L, the change in government purchases also does not shift the labor-supply curve. Since neither curve shifts, the real wage rate, w/P, does not change.

New results come when we consider the determination of consumption and investment. We have that real GDP, Y, is unchanged; real government purchases, G, are higher in year 1 by one unit; and consumption, C, is lower but by much less than one unit.

Consider the expression for real GDP:

$$(12.11) \qquad Y = C + I + G.$$

We have that, in year 1, C declines but by not nearly as much as the rise in G. Consequently, gross investment, I, must fall. In fact, since the decrease in C is relatively small, the decline in I is large. That is, year 1's extra G now crowds out I more than C. In contrast, when the change in G was permanent, we predicted that most or all of the crowding out would be on C.

Effects on the term structure of interest rates

We found that a temporary increase in government purchases, G, did not affect the real interest rate, r. We also found that investment, I, declined. Over time, the decline in investment means that the stock of capital, K, will be lower. A decrease in K reduces the supply of capital services and leads, thereby, to an increase in the market-clearing real rental price, R/P. The rise in R/P leads to an increase in r, in accordance with equation (12.10). Hence, although the current real interest rate does not change, future real interest rates rise.

In our model, the real interest rate, r, corresponds to a short-term real rate. In the real world, bonds are traded with varying maturities. For example, if we think about the indexed U.S. Treasury bonds that we studied in chapter 11, a one-year bond might pay the real rate of return $r(1)$, a five-year bond the real rate of return $r(5)$, and so on. The **term structure of real interest rates** is the relation between the real rate of return, $r(j)$, and the maturity, j. If r(j) increases with j, the term structure is upward sloping; otherwise it is downward sloping.

If we consider, say, a five-year horizon, an individual can hold to maturity a five-year indexed U.S. government bond or can instead hold a sequence of five one-year indexed bonds. In the first case, the real rate of return is *r(5)*. In the second, the real rate of return is an average of the five one-year returns, *r(1)*. Competition in the financial markets will work to equate the anticipated rates of return from the two options (as well as for other options). Hence, *r(5)* will have to be an average of the *r(1)'s* expected to prevail over the next five years.

In the case of a temporary increase in government purchases, short-term real interest rates, such as *r(1)*, did not change initially. However, anticipated future values of *r(1)* increased. Therefore, the average of the *r(1)'s* expected to prevail over the next five years rose. Since *r(5)* equals the average of the expected *r(1)'s*, it follows that *r(5)* rises immediately when government purchases increase. In other words, the model predicts an effect on the term structure of real interest rates. Short-term rates do not change immediately, but longer-term rates increase. Hence, the term structure becomes more upward sloping.

B. Government purchases and real GDP during wartime: empirical

We now evaluate the model's predictions by considering the response of the economy to the temporary changes in government purchases that accompanied U.S. wars. Table 12.2 considers World War I, World War II, the Korean War, and the Vietnam War. We can measure the temporary part of real defense expenditure by the difference between actual spending and an estimated trend. The trend is calculated in our usual manner by fitting a line through the historical data. We focus on the war years in which the estimates of temporary real defense expenditure were at their peaks: 1918, 1943-44, 1952-53, and 1967-68. The values of temporary spending, in 1996 dollars, were $84 billion or 16% of trend real GDP in 1918, $537 billion or 44% of trend real GDP in 1943-44, $56 billion or 3% of trend real GDP in 1952-53, and $46 billion or 1.4% of trend real GDP in 1967-68. Based on these numbers, we can be pretty confident that wartime spending, and possibly other effects from war, were the major influences on the economy during World Wars I and II. That is, we do not have to worry about holding constant other factors. Wartime expenditure would also be a major influence during the Korean War, though not necessarily the dominant force. In the Vietnam War, the temporary military expenditure of only 1.4% of real GDP was unlikely to be the overriding factor, that is, other disturbances were likely to be of comparable or greater significance.

Other U.S. wartime experiences since 1889—the period where we have reliable data—were of lesser magnitude, as gauged by the ratio of temporary defense expenditure to GDP.[7] The next largest ratios were 0.6% during the Spanish-American War in 1898 and 0.5% for the Afghanistan-Iraq conflicts in 2002-03.[8]

The results are broadly similar for World Wars I and II and the Korean War. In each case, real GDP was above its trend but by less than the excess of real defense

---

[7] Lack of data prevents our studying some earlier large wars, notably the U.S. Civil War and the Revolutionary War. These wars, fought on U.S. soil, would bring in important negative effects on real GDP from the destruction of domestic capital stock.

[8] The ratio was 0.6% at the peak of the Reagan defense buildup in 1987. However, this defense spending was not associated with a war and was probably not viewed as temporary. The ratio equaled 0.2% during the Gulf War in 1991.

expenditure from its trend. For example, in 1943-44, the excess for real defense spending of $537 billion was matched by an excess for real GDP of $433 billion. For 1918, the parallel numbers were $84 billion and $42 billion, whereas for 1952-53, they were $56 billion and $49 billion.

Since real GDP was up by less than defense spending, the other components of GDP had to be below trend overall. In 1943-44, the shortfalls from trend were $58 billion in gross investment and $20 billion in non-defense forms of government consumption and investment. Consumption was about equal to trend, and net exports of goods and services were $23 billion below trend.[9] We study net exports in chapter 17.

In 1918, the pattern for the components of GDP was somewhat different—the shortfalls from trend were $21 billion each for gross investment and consumption. Non-defense government consumption and investment and net exports were about equal to trend. In 1952-53, gross investment and consumption were each about equal to trend, non-defense government consumption was $5 billion above trend, and net exports were $11 billion below trend.

The numbers for the Vietnam War were different. In this case, the excess of real GDP from its trend for 1967-68 was $81 billion, greater than the $46 billion excess for real defense spending. Since the excess for GDP exceeded that for defense spending, the other components of GDP had to be above trend overall. The breakdown was $31 billion for consumption and $5 billion each for gross investment and non-defense government purchases. Net exports were $5 billion below trend. The boom in real GDP in 1967-68 likely reflected factors in addition to the extra wartime spending, which was only 1.4% of

[9] The $23 billion breaks down into a $16 billion shortfall of exports from trend and a $7 billion excess of imports from trend.

trend GDP. This view fits with the observation that the estimated excess of real GDP from its trend in 1966—$114 billion—was even greater than that for 1967-68. That is, the economy was experiencing a boom in the mid 1960s before the main increase in defense spending for the Vietnam War. Thus, a reasonable interpretation is that defense spending was not the principal determinant of the excess of real GDP from trend in 1967-68. Therefore, we should focus on the three larger wars to isolate the economic effects of wartime spending.

C. Wartime effects on the economy

The main failing of the model—and quite a striking one—is its prediction that real GDP would be constant during wartime. We got this prediction from our assumption that labor input, L, was fixed. To evaluate this assumption, we look first at the numbers on employment during wartime.

1. Employment during wartime. The number of persons in the military soared during World War II—in 1943-44, military personnel reached 7.7 million persons above its estimated trend. Civilian employment changed relatively little, rising by 1.5 million or 3% above its estimated trend. Putting the two components together, we get that total employment—civilian plus military—was 9.1 million or 17% above trend. The employment patterns for World War I and the Korean War were similar. In 1918, military personnel was 2.5 million above trend, whereas civilian employment was 0.5 million or 1% above trend. In 1952-53, military personnel was 0.7 million above trend, and civilian employment was 0.2 million or 0% above trend. The basic pattern is that the military took in a significant number of persons—in these wartime cases, primarily by the military draft—and total employment expanded by a little more. To do better with our predictions, we have to explain why the total quantity of labor supplied increased so much.

2. Effects of war on labor supply. At this point, there is no settled view among economists about the best way to understand labor supply during wartime. Thus, we consider a number of possibilities and will not be able to reach a definitive explanation. Here are some ideas that have been advanced.

- A large expansion of government purchases, *G*, means that households have less income available after taxes. The negative income effect predicts reductions in consumption and leisure and, hence, an increase in labor supply, L^s. Several considerations influence the size of the income effect. For one thing, we have stressed that the increase in *G* is likely to be temporary, at least if the wars are expected to last no more than a few years and if no major destruction of capital stock and population is anticipated.[10] This consideration makes the income effects on consumption and leisure relatively small. On the other hand, military outlays would not substitute for private consumption in the provision of utility—that is, the parameter λ introduced in the box in section VI.A would be zero. This consideration makes the income effects large in comparison with non-military components of government purchases. Overall, the income effect predicts an

[10] In terms of destruction of human life, the highest U.S. casualty rates in relation to the population were during the Civil War, when roughly 500,000 people died. U.S. war-related deaths during World Wars I and II were 117,000 and 405,000, respectively, well below 1% of the total labor force. Casualties for the Korean and Vietnam Wars were much smaller. For several European countries during World Wars I and II, casualty rates were far higher than anything ever experienced by the United States.

increase in labor supply, L^s. The problem, however, is that the same argument predicts a decrease in consumption, *C*. This prediction conflicts with the finding in Table 12.2 that *C* did not decrease much during the major wars.

- Casey Mulligan (1998) argues that labor supply, L^s, increases during wartime because of patriotism. That is, for a given real wage rate, *w/P*, and for given total real income, people are willing to work more as part of the war effort. The attraction of this argument is that it does not rely on a negative income effect and can, therefore, explain why consumption does not fall much during wartime. However, it may be that the effect of patriotism is more important during a popular war, such as World War II, than in other conflicts.
- From the standpoint of families, we want to understand how the military draft's forced removal of many men would influence the labor supply of those not drafted, especially women. One part of this analysis involves married couples, with the man drafted into the military. In these cases, the postponement of having children would be an important part of the story. Thus, women might participate more in the labor force as a temporary alternative to raising a family or having a larger family. Another consideration involves the postponement of marriage. Through this channel, the military draft would affect the labor supply of single women. That is, women who would otherwise have married and possibly had children found market work to be an attractive, temporary alternative.

The upshot of these arguments is that wartime likely entails an increase in labor supply, L^s. Thus, we now assume that the occurrence of a war raises L^s. We show the effects on the labor market in Figure 12.6. Unlike in Figure 12.4, we now allow for a positive effect of the real wage rate, *w/P*, on L^s. Thus, before the war occurs, the labor-supply curve, shown in blue, slopes upward versus *w/P*. The war shifts the curve rightward to the green one, denoted $(L^s)'$. At any *w/P*, the quantity of labor supplied is larger along the green curve than along the blue one.

The labor-demand curve, denoted by L^d and shown in red, slopes downward versus *w/P*. This curve is the same as the one in Figure 12.4. We still assume that the occurrence of a war does not shift the labor-demand curve.

Before the war, the labor-market clears at the quantity of labor, L^*, and real wage rate, $(w/P)^*$, shown in Figure 12.6. After the war occurs, the quantity of labor rises to $(L^*)'$, and the real wage rate falls to $[(w/P)^*]'$. Thus, when we allow for an increase in labor supply, the model can explain a rise in total employment, as observed in Table 12.2.

3. Effects of war on the real wage rate. Figure 12.6 implies that a war would reduce the real wage rate, *w/P*. This proposition receives a mixed verdict from the main U.S. wartime experiences. If we compute the average percentage deviation of the real wage rate from its trend during the years of the main wars, we get the following:[11]

- World War I (1917-18): -4.0%
- World War II (1942-45): +3.1%

[11] The nominal wage rate is the average hourly earnings of production workers in manufacturing. The real wage rate is the nominal wage rate divided by the GDP deflator. The trend for the real wage rate is calculated in our usual manner. The results for the Korean War come from quarterly data since 1947. The results for World Wars I and II come from annual data for 1889 to 2002.

- Korean War (1951-53): 0.0%

Thus, the predicted negative effect on *w/P* shows up only for World War I. The excess of *w/P* from trend during World War II conflicts with our prediction. For the Korean War, *w/P* deviates negligibly from its trend.

A further analysis suggests that the model might be doing better than these numbers indicate. Price controls and rationing of goods were imposed during World War II and, to a lesser extent, during the Korean War. Consequently, the reported price level, *P*, understated the true price level—typically, households could not buy additional goods just by paying the stated price. For example, to buy more goods, a household might have to pay the black-market price, which exceeded the stated price at a time of price controls and rationing. Since *P* was understated, the measured real wage rate, *w/P*, overstated the true real wage rate. That is, because of rationing, households could not buy *w/P* additional goods with an additional hour of labor. In principle, we could adjust *P* upward (by an unknown amount) to calculate the true price level—that is, the amount that a household would actually have to pay on the black market to buy more goods. The upward adjustment in *P* means that the adjusted real wage rate would be lower than the measured one during World War II and the Korean War. With this amendment, the model would work better. That is, the adjusted real wage rate may have fallen below trend during World War II and the Korean War.

4. Effects of war on the rental market.   We learned from Figure 12.6 that a wartime increase in labor supply, L^s, led to an increase in labor input, *L*. This change affects the rental market, because the rise in *L* tends to increase the marginal product of capital services, MPK. Before the war, the demand for capital services, $(\kappa K)^d$, slopes downward versus the real rental price, R/P, as shown by the blue curve in Figure 12.7. The occurrence of a war shifts this demand rightward to the curve $[(\kappa K)^d]'$, shown in green. We still have that war does not shift the supply curve for capital services, $(\kappa K)^s$, shown in red.

We see from Figure 12.7 that the real rental price, R/P, and the quantity of capital services, κK, increase. For a given capital stock, K, the rise in κK corresponds to an increase in the capital utilization rate, κ. Recall that the real interest rate is given by

(12.10) $$r = (R/P)\cdot\kappa - \delta(\kappa).$$

The increases in R/P and κ imply that r increases. Therefore, we have two new predictions about wartime. The capital utilization rate, κ, and the real interest rate, r, increase.

We looked before at the data on capacity utilization rates. These numbers show that the utilization rate, κ, rose significantly above trend during the Korean War—the average for 1952-53 was 0.025. These data are unavailable before 1948, but another series reveals a sharp rise in capital utilization rates in manufacturing during World War II.[12] Thus, the model's prediction for higher capital utilization during wartime accords with the facts.

In contrast, the predictions for higher real interest rates during wartime seem to conflict with the U.S. data. During the Korean War (1951-53), the real interest rate on 3-month U.S. Treasury Bills was, on average, about equal to its trend value. During World War II (1942-45), the nominal 3-month T-Bill rates were extremely low—less

[12] This series, constructed by the Federal Reserve and the Bureau of Economic Analysis, is for manufacturing production per unit of installed equipment. These data start in 1925 and are therefore unavailable for World War I.

than 1%—and inflation rates averaged 5%. Therefore, real interest rates were negative.[13] In World War I, nominal 3-month T-Bill rates rose from 3% in 1916 to 6% in 1918, but the inflation rate soared—to as high as 16%. Hence, real interest rates were again negative. These occurrences of extremely low real interest rates during wartime are not well understood. Moreover, these puzzling results for U.S. wars contrast with those for the United Kingdom over the history of the many wars fought during the 1700s and up to the end of the Napoleonic Wars in 1814. In those cases, increases in wartime spending were accompanied by rises in real interest rates.[14] That is, with respect to real interest rates, the model accords with the long-term British data but not with the findings for the United States during World Wars I and II.

VIII. Summing Up

We extended the model to include government purchases of goods and services. These purchases were added to the government's budget constraint and were financed by lump-sum taxes net of lump-sum transfers.

A permanent rise in government purchases leads to a long-lasting increase in taxes net of transfers. The strong income effect causes a roughly one-to-one decline in consumption. However, the model predicts little response of real GDP, investment, and the real interest rate. We verified the prediction that cyclical movements in government purchases would have little relation to the cyclical variations in real GDP.

[13] The 5% average inflation rate is based on the reported consumer price index. Because of price controls, the true inflation rates were probably higher during the war, so that the true real interest rates were even more negative than the measured ones.
[14] For studies of the long-term British experience, see Benjamin and Kochin (1984) and Barro (1987).

A temporary rise in government purchases has only a weak income effect. Therefore, consumption reacts relatively little, and investment falls substantially. We checked out these predictions for the temporary changes in purchases associated with major U.S. wars. The evidence brought out a number of unresolved puzzles involving large wartime expansions of total employment (including the military) and the failure of real interest rates to rise systematically during wartime. Some of the observations could be explained from positive effects of patriotism on wartime labor supply and from effects of wartime rationing and price controls.

Questions and Problems

Mainly for review

Problems for discussion

12.x. Government consumption in the national accounts

The national accounts treat all of government consumption, G, as part of real GDP. But suppose that the public services derived from government consumption are an input to private production, say

$$Y = F(\kappa K, L, G).$$

In this case, public services are an intermediate product—a good that enters into a later stage of production. Hence, we ought not to include these services twice in real GDP—once when the government buys them and again when the services contribute to private production.

a. Suppose that businesses initially hire private guards. Then subsequently the government provides free police protection, which substitutes for the private guards. Assume that the private guards and public police are equally efficient and receive the same wage rates. How does the switch from private to public protection affect measured real GDP?

b. How would you change the national accounts to get a more accurate treatment of government consumption? Is your proposal practical? (These issued are discussed by Simon Kuznets [1948, pp. 156-57] and Richard Musgrave [1959, pp. 186-88].)

12.x. Public ownership of capital and the national accounts

Until the revision of 1996, the U.S. national accounts included in GDP the government's purchases of goods and services from the private sector. The accounts neglected any contribution to output from the flow of services on government owned capital. The accounts also did not subtract depreciation of this capital to calculate net domestic product.

a. In the pre-1996 system, what happened to GDP if the government gave its capital to a private business and then bought the final goods from that business?

b. In the current system of national accounts, the GDP includes an estimate of the flow of income generated by public capital. However, this income flow is assumed to equal the estimated depreciation of the public capital. Redo part a. in the context of this system.

12.x. A prospective change in government consumption

Suppose that people learn in the current year that government consumption, G_t, will increase in some future year. Current government consumption, G_1, does not change.

a. What happens in the current year to real GDP, Y, consumption, C, and investment, I?

b. Can you think of some real-world cases to which this question applies?

12.x. The price level during the Korean War

In 1949, the inflation rate was negative. With the start of the Korean War, the price level (GDP deflator) rose at an annual rate of 10% from the second quarter of 1950 to the first quarter of 1951. In contrast, the inflation rate was only around 2% from the first quarter of 1951 to the first quarter of 1953. The table shows for various periods the inflation rate, π, the growth rates, μ, of currency and M1, and the growth rate of real government consumption and investment, $\Delta G/G$. Can we use these data to explain the surge of the price level at the start of the Korean War, followed by a moderate inflation rate?

(This question does not have a definite answer. However, a significant fact is that price controls were stringent during World War II. People may have expected a return to these controls when the Korean War started in June 1950. The controls implemented as of December 1950 were more moderate than those applied during World War II.)

Period	**Inflation rate (π)**	**M1 growth rate (μ)**	**Currency growth rate (μ)**	**Growth rate of G**
	(all figures in percent per year)			
1949.1 to 1950.2	-1.2	1.8	-1.6	2.5
Start of war				
1950.2 to 1951.1	10.2	4.0	-0.5	19.3
1951.1 to 1952.1	1.7	5.3	4.7	32.0
1952.1 to 1953.1	2.2	3.2	4.5	9.4

12.x. The role of public services

We studied in the box on p. xxx the role of public services in providing utility to households. We assumed that each unit of government consumption, G, was equivalent to λ units of private consumption, C, in terms of household utility.

a. Consider various categories of government expenditure, such as military spending, police, highways, public transit, and research and development. How do you think the coefficient λ varies across these categories?

b. Suppose that G rises permanently by one unit. What are the responses of real GDP, Y, consumption, C, and investment, I? How do the results depend on the size of the coefficient λ?

Table 12.1			
General Government Expenditure as a Ratio to GDP:			
Averages by Country for 2000-2002			
Country	**Spending ratio**	**Country**	**Spending ratio**
Australia	0.33	Latvia	0.39
Austria	0.50	Lithuania	0.32
Azerbaijan	0.23	Luxembourg	0.41
Belgium	0.47	Malaysia	0.27
Bolivia	0.37	Moldova	0.24
Botswana	0.42	Netherlands	0.42
Brazil	0.37	New Zealand	0.38
Bulgaria	0.41	Norway	0.42
Canada	0.38	Panama	0.28
Chile	0.23	Peru	0.29
Croatia	0.51	Russia	0.16
Czech Rep.	0.32	Singapore	0.18
Denmark	0.52	Slovakia	0.39
Ecuador	0.25	Slovenia	0.42
Egypt	0.26	South Korea	0.24
Estonia	0.37	Spain	0.38
Finland	0.44	Sweden	0.53
France	0.49	Switzerland	0.38
Germany	0.47	Taiwan	0.25
Greece	0.47	Thailand	0.18
Hong Kong	0.13	Trinidad	0.32
Hungary	0.53	Ukraine	0.26
Ireland	0.30	United Kingdom	0.38
Italy	0.46	United States	0.32
Japan	0.37	Venezuela	0.23
Kazakhstan	0.22		

Note: The table shows the ratio of general government expenditure to GDP, averaged for 2000-2002. Countries are included only when data were available for general government expenditure, which includes spending at all levels of government on goods and services, transfers, and net interest. Data for countries other than the United States are from Economist Intelligence Unit, *EIU Country Data*.

Table 12.2				
U.S. Wartime Spending, GDP, and Employment				
I: Real GDP and Components. Each entry is the deviation from trend in billions of 1996 dollars. Values in parentheses are percentage of own trend.				
	Wartime years			
	1918	**1943-44**	**1952-53**	**1967-68**
Category of GDP:				
Defense spending	84 (679)	537 (317)	56 (25)	46 (15)
% of trend real GDP	16	44	3	1
Real GDP	42 (8)	433 (36)	49 (3)	81 (2)
Consumption	-21 (-5)	-1 (0)	0 (0)	31 (1)
Gross investment	-21 (-28)	-58 (-51)	0 (0)	5 (1)
Non-defense government	0 (0)	-20 (-19)	5 (3)	5 (1)
Net exports	0	-23	-11	-5
II: Employment. Each entry is the deviation from trend in millions. Values in parentheses are percentage of own trend.				
Category of employment:				
Total employment	3.0 (8)	9.1 (17)	0.9 (1)	1.0 (1)
Civilian employment	0.5 (1)	1.5 (3)	0.2 (0)	0.4 (1)
Military personnel	2.5 (566)	7.7 (296)	0.7 (24)	0.6 (19)

Notes: In part I, each cell shows the deviation of an expenditure component from its estimated trend in billions of 1996 dollars. The values in parentheses express the deviations as a percentage of the trend. For example, as an average for 1943 and 1944, defense expenditure was $537 billion or 317% above its own trend, real GDP was $433 billion or 36% above its own trend, and so on. Each expenditure component is the nominal value divided by the deflator for the GDP. (The trend for real GDP was constrained to equal the sum of the trends estimated for the components of GDP.) In part II, total employment is the sum of civilian employment and military personnel. Each entry shows the deviation of a component of employment from its own trend in millions. For example, as an average for 1943 and 1944, total employment was 9.1 million or 17% above its own trend, civilian employment was 1.5 million or 3% above its own trend, and military personnel was 7.7 million or 296% above its own trend. (The trend for total employment was constrained to equal the sum of the trends estimated for its two parts, civilian employment and military personnel.) Data for the last three wars are from Bureau of Economic Analysis. Data for 1918 are from John Kendrick (1961), Christina Romer (1988), and U.S. Department of Commerce (1975).

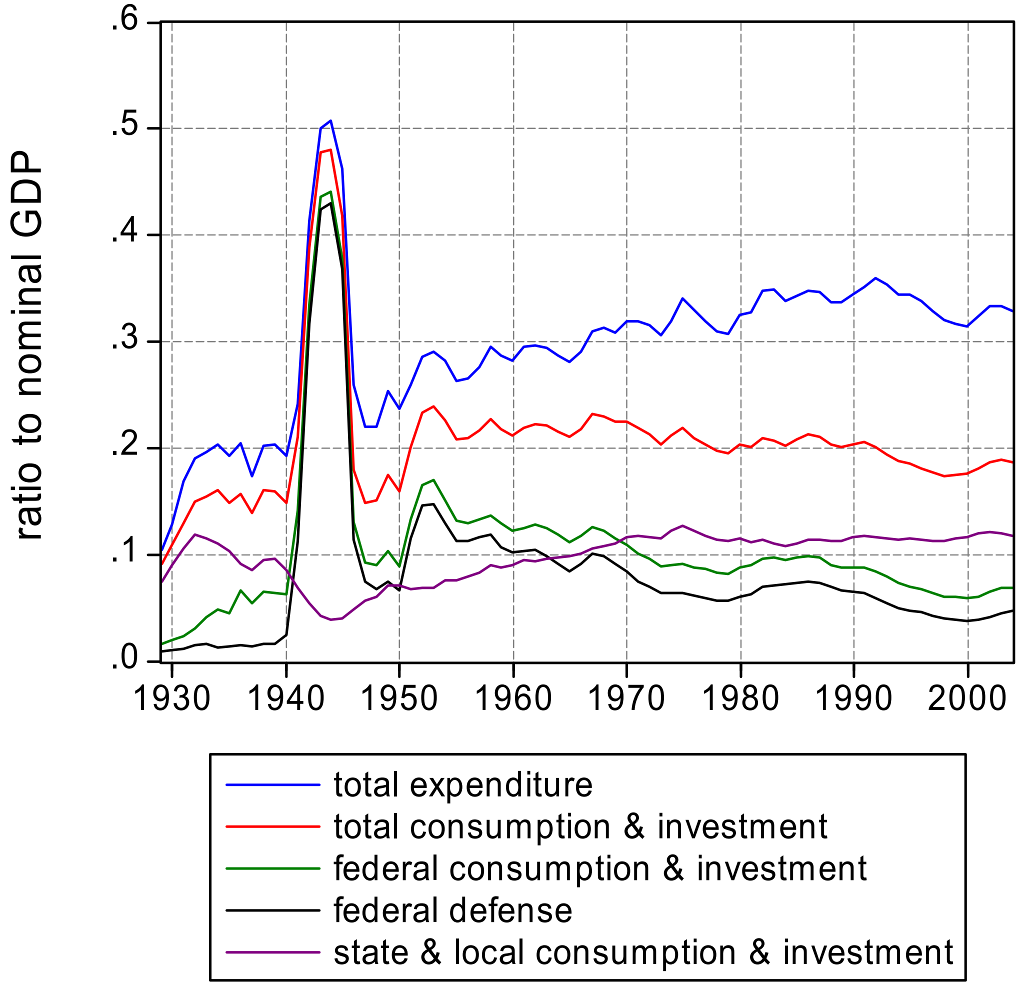


Figure 12.1

Total Government Expenditure and Government Consumption and Investment

The graphs show the ratio of each government expenditure component to nominal GDP.

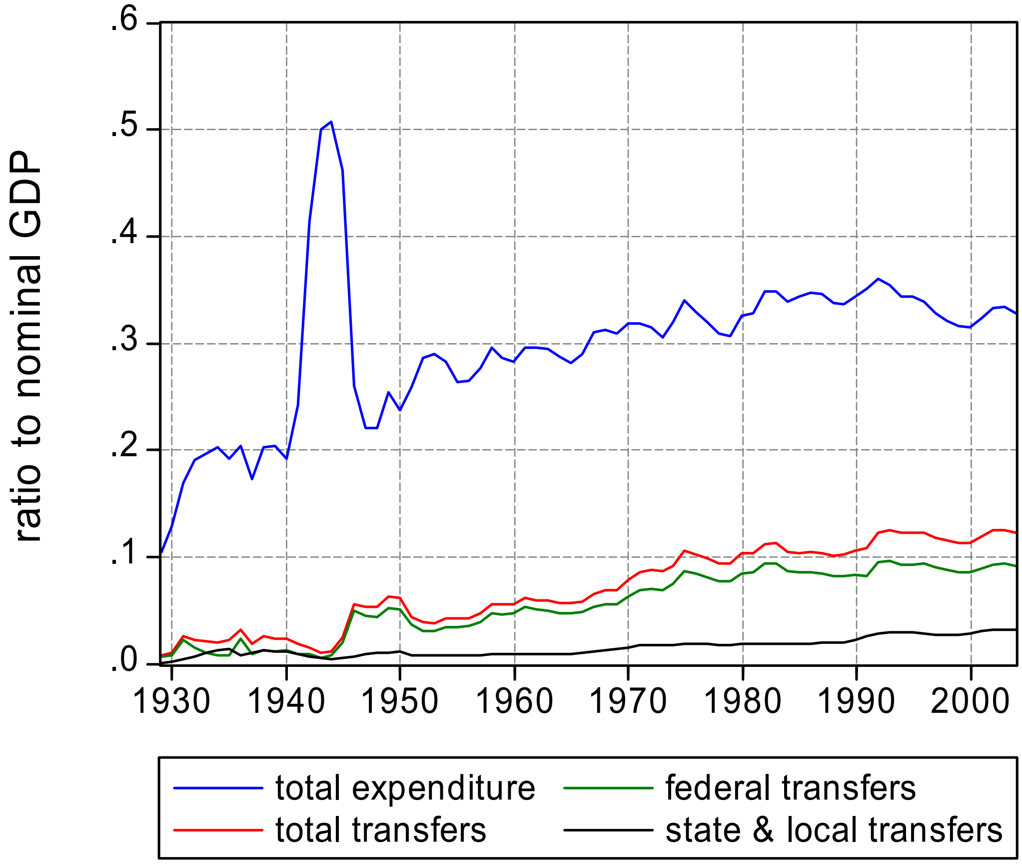


Figure 12.2

Total Government Expenditure and Transfer Payments

The graphs show the ratio of each government expenditure component to nominal GDP.

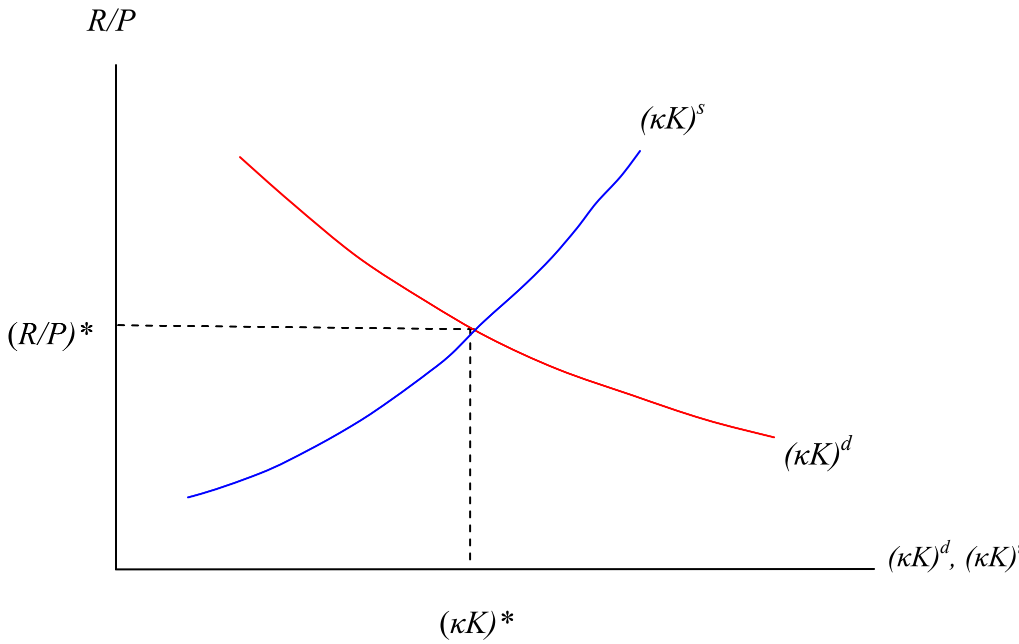


Figure 12.3

Clearing of the Market for Capital Services

This construction comes from Figure 8.8. The demand curve for capital services, $(\kappa K)^d$, comes from the equation of the marginal product of capital services, MPK, to the real rental price, *R/P*. When *R/P* rises, the quantity of capital services demanded falls. The supply of capital services, $(\kappa K)^s$, applies for a given capital stock, *K*. If *R/P* rises, owners of capital raise the capital utilization rate, *κ*. Therefore, the quantity of capital services supplied rises. The market clears where the two curves intersect.

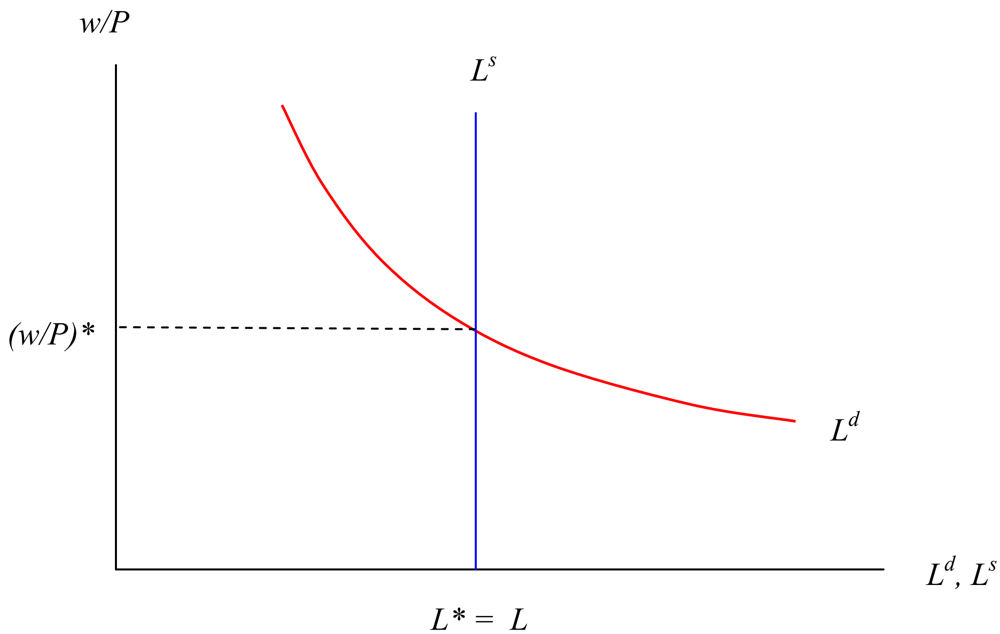


Figure 12.4

Clearing of the Labor Market

This construction comes from Figure 8.3. The demand curve for labor, L^d, shown in red, comes from the equation of the marginal product of labor, MPL, to the real wage rate, *w/P*. When w/P rises, the quantity of labor demanded falls. We assume here that labor supply, L^s, shown in blue, equals the constant *L*. The market clears where *w/P = (w/P)**, so that the quantity of labor demanded equals *L*.

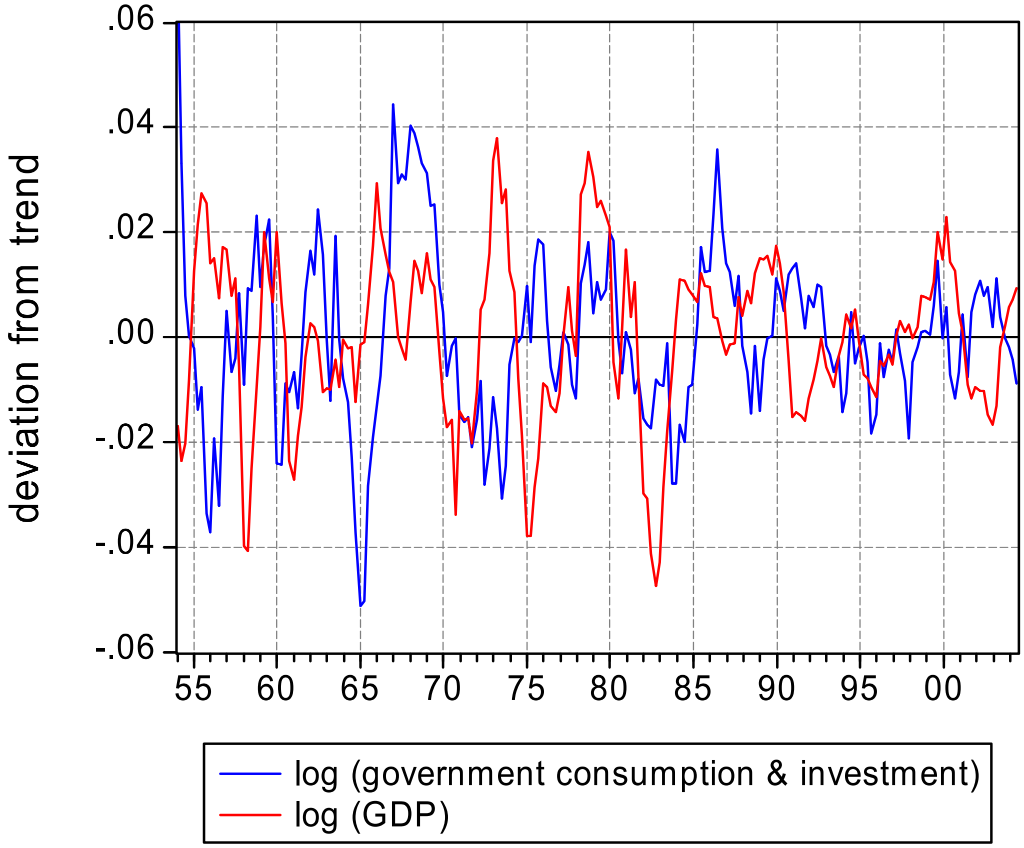


Figure 12.5

The Cyclical Parts of GDP and Government Consumption & Investment

The red graph is the deviation of the log of real GDP from its trend. The blue graph is the deviation of the log of government consumption and investment from its trend. Government consumption & investment is about as variable as GDP but is acyclical, that is, it has close to a zero correlation with GDP.

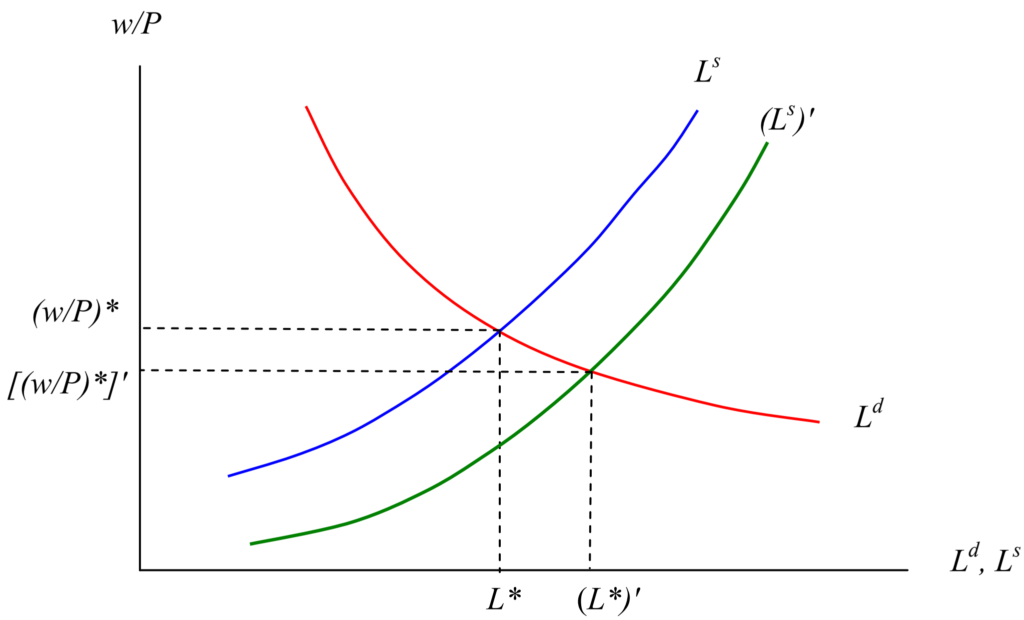


Figure 12.6

Effect of a Wartime Increase in Labor Supply on the Labor Market

The downward-sloping labor-demand curve, L^d, shown in red, comes from Figure 12.4. We now allow the real wage rate, *w/P*, to have a positive effect on labor supply, L^s, shown by the blue curve. We assume that the occurrence of a war shifts the labor-supply curve rightward from L^s to $(L^s)'$, shown in green. Consequently, the quantity of labor input rises from L^* to $(L^*)'$. The real wage rate falls from $(w/P)^*$ to $[(w/P)^*]'$.

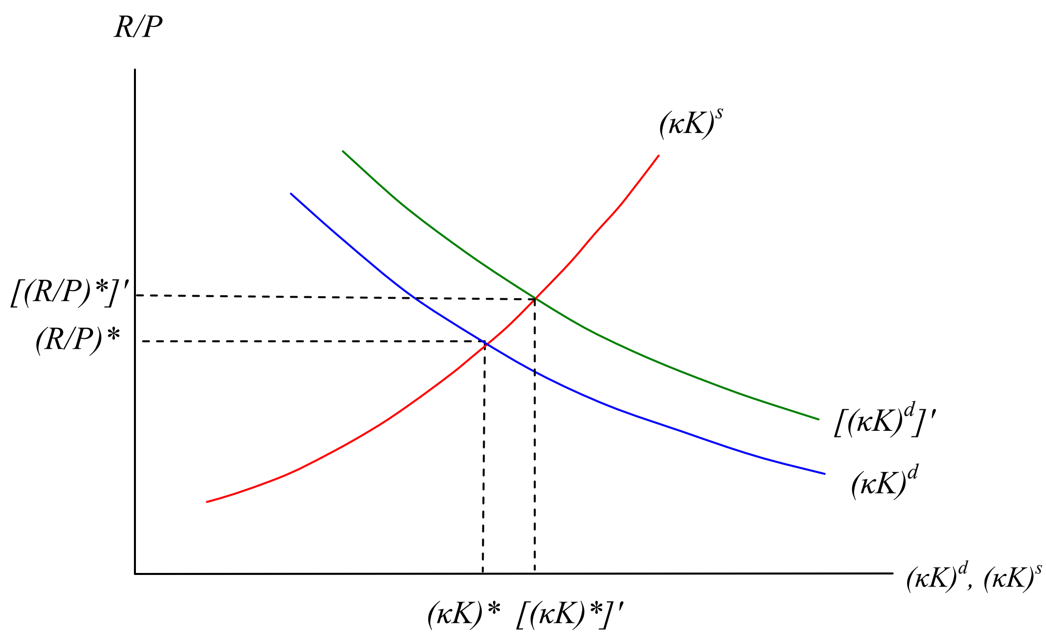


Figure 12.7

Effect of a Wartime Increase in Labor Input on the Market for Capital Services

The increase in employment from L^* to $(L^*)'$, shown in Figure 12.6, raises the schedule for the marginal product of capital, MPK. Therefore, the demand curve for capital services shifts rightward, from $(\kappa K)^d$, shown in blue, to $[(\kappa K)^d]'$, shown in green. Consequently, the market-clearing real rental price of capital rises from $(R/P)^*$ to $[(R/P)^*]'$. The quantity of capital services expands from $(\kappa K)^*$ to $[(\kappa K)^*]'$. This increase in capital services corresponds, for a given capital stock, K, to a rise in the capital utilization rate, κ.

January 12, 2005

Chapter 13

Taxes

In the previous chapter, we took an unrealistic view of the government by assuming lump-sum taxes and transfers. In the model, the amount that a household paid as taxes or received as transfers has nothing to do with the household's income or other characteristics. In the real world, governments levy a variety of taxes and pay out a lot of transfers, but none of them look like the lump-sum taxes and transfers in our model.

Usually, a household's taxes and transfers depend on its actions. This dependence motivates changes in behavior. For example, taxes on labor income discourage households from working and earning income. Transfers to the unemployed motivate people not to be employed. Taxes on asset income discourage saving. Overall, the systems of taxes and transfers create substitution effects that influence labor supply, production, consumption, and investment. In this section, we extend the model to incorporate some of these effects. However, before we extend the theory, it is useful to have an overview of tax collections in the United States.

I. Taxes in the United States

Figure 13.1 shows the breakdown of tax revenue by major type for the U.S. federal government since 1929. Individual income taxes have been a reasonably stable share since the end of World War II. The share was 50% in 2000 and 2001 but dropped to 45% in 2002. The next component—contributions (really taxes) for social-insurance funds—rose from 11% in 1950 to almost 40% by the late 1980s. Since then, the share has been reasonably stable. Another part, corporate profits taxes, fell from about 25% of the total at the end of World War II to around 8% by the early 1980s. Since then, the share has fluctuated between 7% and 11%. The category that includes excise taxes and customs duties fell from 25% of federal receipts at the end of World War II to 6% in 2002. Finally, we have the transfers from the Federal Reserve to the U.S. Treasury—this transfer represents the government's revenue from printing money. This component is relatively minor—it was about 2% of federal receipts during the 1970s and 1980s but fell to about 1% in 2002.

Before World War II, excise taxes and customs duties were relatively much more important. This category comprised about 60% of federal receipts during the Great Depression of 1932-35 and were the major source of federal revenues before World War I. Individual income taxes began in 1913, except for some levies around the Civil War and in 1895. Corporate profits taxes began in 1909. Notice also that the levies for social-insurance programs were small until the beginning of the unemployment insurance program in 1936 and social security in 1937.

Figure 13.2 shows a breakdown for the revenue of state and local governments. Property taxes were traditionally the largest component, but the share fell from 65% in 1929 to less than 40% after World War II. From the late 1970s until 2002, the share was between 20% and 24%. In the pre-World War II period, the relative decline in property taxes corresponded to growth in sales taxes. These levies increased from 6% of total receipts in 1929 to 22% in 1941. Since the end of World War II, this share has stayed between 24% and 29%. More recently, state and local governments turned to individual income taxes. This category rose from 9% of total receipts in 1946 to 20-22% from 1987 to 2002. The other form of state & local revenue that became important after World War II is federal grants-in-aid (transfers from the federal government to state & local governments). These revenues climbed from 8% of the total in 1946 to 23% in the late 1970s. Then, after a fall to 16% in the late 1980s, the share rose back to 23% in 2002. (Federal grants-in-aid go primarily for welfare, medical care, transportation, education, housing, and training programs.)

Figure 13.3 shows the federal share of total government receipts. (Federal grants-in-aid are excluded here from total receipts.) The federal share was 35% in 1929, fell to a low point of 20% in 1932, then rose during the New Deal period to reach 49% by 1940. After a peak of 81% during World War II, the share declined to around 65% by the early 1970s. From then until 2002, the federal share varied between 64% and 69%.

The red graph in Figure 13.4 shows one measure of an overall tax rate—the ratio of total government receipts (exclusive of federal grants-in-aid to state & local governments) to GDP. The ratio rose from 10% in 1929 to 16% in 1945 and 23% in 1945. After a fall to 20% in 1949, the ratio increased to 25% in 1960 and 28% in 1969. The ratio then stayed between 26% and 28% until 1994, when it rose to its all-time high of 31% in 2000. Then the ratio fell to 27% in 2002.

II. Types of Taxes

Some taxes fall on forms of income—individual income taxes, corporate profits taxes, and contributions for social security and Medicare, which are based on wage earnings. Other taxes are based on expenditures—sales taxes, excise taxes, and customs duties. Many countries outside the United States use value-added taxes (VAT), which are like sales taxes, but assessed at various stages of production. Still other forms of taxes are based on ownership of property and are, therefore, a form of wealth tax. An important point is that, for all of these taxes, the amount that someone pays depends on his or her economic activity. None of them look like the lump-sum taxes in our model.

For taxes that fall on income, it is important to distinguish the **marginal tax rate** from the **average tax rate**. The marginal tax rate is the additional tax paid on an additional dollar of income. The average tax rate is the ratio of total taxes paid to total income. The marginal tax rate will turn out to have substitution effects that influence the behavior of households and businesses. The average tax rate will determine the government's revenue. That is, revenue equals the average tax rate multiplied by total income.

An important property of the U.S. federal individual income tax is that the marginal tax rate rises with income. Table 13.1 illustrates the nature of marginal and average income-tax rates with numbers that approximate the U.S. individual income-tax system in 2002. The first $20,000 of income is assumed to be tax exempt (corresponding to the standard deduction of $7850 and personal exemptions of $12,000 for a family of four). Income in excess of $20,000 is taxed at a 10% rate. Therefore, at an income of $20,000, the marginal income-tax rate becomes 0.10. However, the average income-tax rate is still zero, because no taxes have yet been paid. At an income of $32,000, taxes reach $1200, so that the average tax rate is 0.04 (1200/32000). At this point, the marginal income-tax rate rises to 0.15 for the next $35,000 of income. Therefore, at $67,000, the taxes paid are $6450, and the average tax rate is 0.10. Then the marginal income-tax rate rises to 0.27 for the next $66,000 of income, and so on. Notice that all income above $327,000 is taxed at a marginal rate of 0.385. That is, the marginal income-tax rate does not rise with income once income exceeds $327,000.

Table 13.1

The U.S. Graduated-Rate Income Tax in 2002

income level ($)	taxes ($)	marginal tax rate	average tax rate
0	0	0	0
20,000	0	0.10	0
32,000	1200	0.15	0.04
67,000	6450	0.27	0.10
133,000	24,270	0.30	0.18
192,000	41,970	0.35	0.22
327,000	89,220	0.386	0.27
1,000,000	348,998	0.386	0.35

Note: The income-tax rate is 0 for the first $20000 of income, 0.10 for the next $12,000 of income, 0.15 for the next $35,000 of income, 0.27 for the next $66,000 of income, 0.30 for the next $59,000 of income, 0.35 for the next $135,000 of income, and 0.386 on all additional income. This structure is close to the one prevailing in the United States in 2002.

The table illustrates two important points about the U.S. federal income tax. First, the marginal income-tax rate rises with income until income reaches $327,000. That is why the system is said to have a **graduated-rate** structure (sometimes called a progressive tax). However, after $327,000, the marginal tax-rate schedule is **flat**, rather than graduated. Second, the marginal tax rate is always higher than the average tax rate.

That is because the average tax rate incorporates the low taxes paid on the early portions of income, including the zero tax paid on the first $20,000 of income. However, as income becomes very high—for example, above the $1,000,000 shown in the table—the average tax rate approaches the top marginal tax rate of 0.386.

Table 13.1 is a great simplification of the complex individual income tax system that prevails in the United States. For example, the calculations ignore many legal things that households can do to reduce their taxes. These tax-reducing activities including itemized deductions (for mortgage interest payments, charitable contributions, state & local income taxes, and other items) and various credits and adjustments (such as the earned-income tax credit, child-care credit, and contributions to pension funds). Some fringe benefits paid by employers, notably for health insurance and pensions, also escape or defer taxation.

Higher-income taxpayers are more able to deal with the complexity of the income tax and also benefit more from some of the available deductions. For these reasons, economists have sometimes questioned whether the individual income tax is as graduated as it seems from the explicit tax-rate structure. However, Table 13.2 shows that the individual income tax is, in fact, highly progressive in the sense that upper-income persons pay proportionately far more than their share of the taxes. The entries in the table show values from 1970 to 2000 of the shares of the total individual income taxes paid by the top 1%, 5%, 10%, 25%, and 50% of tax returns. For example, in 1970, the highest 1% of tax returns, ranked by taxes paid, accounted for 16.7% of all taxes. The top 50% of the returns paid 83.0% of the taxes. Thus, relatively small numbers of tax returns paid a large portion of the taxes. By 2000, this pattern became even more pronounced. In that year, the top 1% of returns accounted for 37.4% of all taxes, whereas the top 50% had 96.1%. In other words, in 2000, the bottom half of tax returns, ranked by taxes paid, accounted for less than 4% of tax payments.

The numbers in brackets in Table 13.2 show the fractions of a broad concept of income reported on income-tax forms—called **adjusted gross income**—received by the corresponding tax returns. For example, in 1970, the top 1% of returns ranked by taxes paid had 7.4% of the adjusted gross income. However, this group paid 16.7% of the taxes. The excess of the share of taxes paid over the share of income received for upper-income taxpayers is a measure of the progressivity of the income-tax structure. By 2000, the top 1% had an even higher share of income, 20.8%. But this group again paid even

Table 13.2

Shares of Individual Income Taxes Paid by Persons who Paid the Most Taxes

The first number in each cell is the percentage of total taxes paid by the indicated range of tax returns. The number in brackets is the share of adjusted growth income represented by this group of tax returns.

	Range of tax returns, based on taxes paid.				
	Top 1%	**Top 5%**	**Top 10%**	**Top 25%**	**Top 50%**
1970	16.7 [7.4]	31.4 [18.3]	41.8 [28.0]	62.2 [49.4]	83.0 [74.7]
1980	17.4 [7.8]	33.7 [19.2]	45.0 [29.1]	66.6 [51.4]	87.0 [76.7]
1990	25.1 [14.0]	43.6 [27.6]	55.4 [38.8]	77.0 [62.1]	94.2 [85.0]
2000	37.4 [20.8]	56.5 [35.3]	67.3 [46.0]	84.0 [67.2]	96.1 [87.0]

Source: R. Glenn Hubbard (2002, p. 3).

more of the taxes, 37.4%. The table provides similar information for other taxpayer groups. For example, in 2000, the top 50% of returns ranked by taxes paid received 87.0% of the adjusted gross income while paying 96.1% of the income taxes.

Another important form of income tax in the United States, which the government amusingly calls a contribution, is the levy on wage earnings and self-employment income to finance social security and Medicare.[1] This tax is much simpler than the individual income tax. In 2003, covered employees paid 6.2% of earnings up to an earnings ceiling of $87,000 to finance the old-age, survivors, and disability program (OASDI) and 1.45% of all earnings to finance Medicare. Employers paid an equal amount. Thus, the combined marginal income-tax rate was 15.3% for labor earnings between 0 and $87,000 and 2.9% thereafter. For incomes between 0 and $87,000, the average tax rate equals the marginal tax rate. This characteristic applies to **flat-rate tax systems**. However, at $87,000, the system has a single, sharp reduction in the marginal tax rate to 2.9%. As income rises above $87,000, the average tax rate falls gradually from 15.3% to 2.9%. Therefore, in this range, the marginal tax rate is less than the average tax rate. This pattern is opposite to the one found in the individual income tax.

Empirically, it is difficult to measure marginal income-tax rates because these rates differ across persons and types of income. The blue graph in Figure 13.4 gives an estimate of average marginal income-tax rates in the United States.[2] The measure considers the two most important forms of income taxes—the federal individual income

[1] These levies are more like taxes than contributions because the services that individuals get do not depend very much on the amount that an individual pays. In the case of social security, the benefits do depend somewhat on the amounts paid in over one's lifetime. Hence, the payments are partly a tax and partly a contribution.
[2] The construction weights households by their adjusted gross incomes. That is, higher-income households count more in the calculation of the average.

tax and social-security taxes. These two levies accounted for 85% of total federal receipts in 2002. Because of data limitations, the measure does not include state & local income taxes.

Figure 13.4 shows that the average marginal income-tax rate was very low in the pre-World War II years. The reason is that individual income taxes covered a small minority of the population and because social-security taxes were not yet significant. The tax rate rose dramatically to 26% at the end of World War II in 1945. Then the rate fell to 18% in 1949 before increasing back to 26% during the Korean War in 1952-53. From then until the early 1960s, the rate remained between 24% and 26%. Then the rate fell to 23% in 1965 because of the Kennedy-Johnson income-tax cuts. The rate then rose gradually to its peak of 39% in 1981. Subsequently, the rate fell because of the Reagan tax cuts to reach 30% in 1988. The rate then rose to 31% in 1990 and 32% in 1994 because of the tax-rate increases under the first President Bush and Clinton.[3] However, these increases in marginal income-tax rates were much smaller than the ones that occurred from 1965 to 1981.

III. Taxes in the Model

To incorporate tax rates into the model, we can start with the household budget constraint from equation (12.4). When written without year subscripts, the constraint is

$$(13.1) \qquad C + (1/P) \bullet \Delta B + \Delta K = (w/P) \cdot L^s + r \cdot (B/P + K) + V/P - T/P.$$

Up to now, we regarded transfers and taxes as lump sum. Therefore, a household's real transfers net of real taxes, $V/P - T/P$, did not depend on the household's characteristics,

[3] At this writing, the series has not been calculated after 1994.

including its income. Now we want to allow a household's transfers and taxes to depend on these characteristics. To simplify, we focus on the determinants of real taxes, T/P.

The various taxes that exist in the United States and other countries can be represented as levies on the terms that appear on the two sides of equation (13.1). Sales, excise, and value-added taxes depend on consumption, C. Labor-income taxes—for example, from the individual income tax and the social-security payroll tax—depend on real wage income, $(w/P)\cdot L^s$. Taxes on asset income, a part of the individual income tax, depend on the term $r\cdot(B/P + K)$.[4] The income base for this tax in the real world includes interest, dividends, and capital gains.[5]

We now assess the economic effects from taxation. In order to affect real GDP, a tax has to influence the quantities of one of the factors of production, labor and capital services. Therefore, the various taxes break down into whether they affect labor or capital services or both. We get the main results by considering two types of taxes—one that depends on labor income and another that depends on asset income.

A. A tax on labor income

We start with a tax on labor income, such as the individual income tax or the payroll tax that finances social security. Let τ_w be the marginal tax rate on labor income.

[4]U.S. taxes are levied on interest payments computed from the nominal interest rate, $i\cdot(B/P)$, rather than the real interest rate, $r\cdot(B/P)$. This treatment of interest income leads to an effect of the inflation rate, π, on real taxes. Another real-world complication is that only parts of household interest expenses are allowed as deductions from income for tax purposes. In the United States, the deduction applies to itemized deductions for interest on home mortgages and debt used to purchase financial assets.

[5]To consider a tax on corporate profit, we could reintroduce real business profit, Π, as a form of household income. We dropped Π before because it equaled zero in equilibrium. However, in most tax systems, the definition of profit differs from the one in our model. The most important difference is that our definition includes the real rental payments to capital, $(R/P)\cdot K$, as a negative term. In the real world, only parts of this rental income—depreciation and interest expenses—are allowed as deductions from income in the computation of corporate profits taxes. With this definition of profit, the corporate profits tax amounts to another levy on the income from capital. Since the income from capital is also taxed at the household level, the corporate profits tax is often described appropriately as **double taxation** of income from capital.

To simplify, we do not allow for a graduated-rate structure for τ_w, as in the U.S. individual income-tax system. Rather, we assume that τ_w is the same at all levels of income. However, households may be allowed deductions and credits that reduce the amount of taxes paid.[6] We also assume that τ_w does not change over time—at least, households do not anticipate that future tax rates will differ from the current rate.[7]

The amount of revenue collected by a tax on labor income equals the average tax rate multiplied by the amount of labor income. The average tax rate depends on the marginal tax rate, τ_w, and on other features of the tax system. For example, an increase in deductions and credits would lower the average tax rate for a given τ_w. However, for a given system of deductions and credits, an increase in τ_w typically goes along with an increase in the average tax rate. Therefore, a rise in τ_w leads to more tax revenue unless the amount of labor income falls sharply.

To assess the economic effects from a tax on labor income, we have to extend our previous analysis of household labor supply. The important thing to reconsider is the substitution effect between leisure and consumption. Without taxation of labor income, this substitution effect depended on the real wage rate, w/P. If a household raised labor supply by one unit, it raised real wage income, $(w/P)\cdot L^s$, by w/P units. This extra income enabled the household to increase consumption by w/P units. At the same time, the rise in labor supply by one unit meant that leisure fell by one unit. Therefore, the household could substitute w/P units of consumption for one unit of leisure. If w/P rose, this deal

[6] In some cases, a household's tax payment could be negative, that is, households receive a transfer from the government. In the U.S. system, these negative taxes arise because of "refundable" tax credits, that is, credits that not only reduce taxes but also allow for cash payments when the computed taxes are negative. The most important of these refundable credits is the earned-income tax credit (or EITC).

[7] Such differences would motivate households to work more in years with relatively low tax rates and less in years with relatively high tax rates. That is, anticipated changes in τ_w over time have intertemporal-substitution effects on labor supply.

became more favorable. Hence, we predicted that households would raise the quantity of labor supplied, enjoy less leisure, and consume more.

The new consideration is that an extra unit of labor income is now taxed at the marginal income-tax rate, τ_w. If the household raises labor supply by one unit, it again raises pre-tax real labor income, $(w/P)\cdot L^s$, by w/P units. This extra income enters on the right-hand side of the budget constraint in equation (13.1). However, the additional labor income raises the household's real taxes, T/P, by τ_w units. These extra taxes also appear on the right-hand side of the equation, but with a negative sign. Overall, the right-hand side rises by $(1-\tau_w)\cdot(w/P)$ units. That is, after-tax real labor income rises by this amount. We also see from the equation that the household can use this additional after-tax real income to increase consumption, C, on the left-hand side by $(1-\tau_w)\cdot(w/P)$ units.

We found that, by working one more unit, the household can raise consumption, C, by $(1-\tau_w)\cdot(w/P)$ units. The increase in work by one unit still means that leisure falls by one unit. Therefore, the deal is that households can substitute $(1-\tau_w)\cdot(w/P)$ units of C for one unit of leisure. Another way to say this is that, in the presence of a labor-income tax, the substitution effect on labor supply depends on the **after-tax real wage rate**, $(1-\tau_w)\cdot(w/P)$, rather than the pre-tax real wage rate, w/P. If the marginal tax rate, τ_w, rises, for a given w/P, $(1-\tau_w)\cdot(w/P)$ falls. Hence, we predict that households would reduce the quantity of labor supplied, take more leisure, and consume less.

We stressed before that labor supply depends also on income effects. The question is, what income effects arise when the marginal income-tax rate, τ_w, rises? We can see from equation (13.1) that a household's real income on the right-hand side still depends on real transfers net of real taxes, $V/P - T/P$. Recall also from chapter 12 that the government's budget constraint requires

$$(12.8) \qquad V/P - T/P = -G.$$

Therefore, if we are holding fixed the government's purchases of goods and services, G, we must also be holding fixed real transfers net of real taxes, $V/P - T/P$. Hence, for given G, we do not get any changes in household real income through the term $V/P - T/P$. In other words, if G is fixed, there are no income effects.

We need to explore this result further, because it seems that a rise in the marginal income-tax rate, τ_w, would have a negative income effect. Suppose, however, that we are holding fixed the government's purchases, G. Since the government has to balance its budget (because we do not yet allow for government borrowing), real transfers net of real taxes, $V/P - T/P$, must be unchanged to satisfy the government's budget constraint. Therefore, when τ_w increases, something else must be happening to keep $V/P - T/P$ from changing.

One possibility is that the government adjusts other features of the tax system to keep the total real taxes collected, T/P, unchanged. For example, marginal income-tax rates, τ_w, might rise in the individual income tax but deductions and credits also rise so as to keep T/P constant. Another possibility is that the government is shifting from collecting revenue through a tax that has a relatively low marginal tax rate—for example, the social-security payroll tax—to one that has a relatively high marginal rate—such as the individual income tax. Shifts of this kind would raise the overall marginal tax rate on labor income, τ_w, for a given total of real taxes collected, T/P.

Another possibility is that real tax revenue, *T/P*, rises along with the increase in τ_w, and all of the extra revenue pays for added real transfers, *V/P*. In that case, the term *V/P - T/P* is again unchanged.

Finally, we could have that real tax revenue, *T/P*, rises along with the increase in τ_w, and the extra revenue pays for added government purchases, *G*. In this case, the economic effects are a combination of two things: the rise in τ_w, which we are now studying, and the rise in *G*, which we considered in chapter 12. In order to keep things straight, it is best to analyze the two effects separately. Thus, we are now assessing the effect from an increase in τ_w for a given *G*. In this case, there are no income effects on labor supply. We can therefore be confident that, for a given real wage rate, *w/P*, the rise in τ_w reduces the quantity of labor supplied, L^s, through the substitution effect from a lower after-tax real wage rate, $(1-\tau_w)\cdot(w/P)$.

Figure 13.5 shows the effects on the labor market from an increase in the marginal tax rate on labor income, τ_w. As in Figure 12.6, we plot the pre-tax real wage rate, *w/P*, on the vertical axis. For a given *w/P*, a higher τ_w implies a lower after-tax real wage rate, $(1-\tau_w)\cdot(w/P)$. Therefore, a rise in τ_w shifts the labor supply curve leftward from the blue one labeled L^s to the green one labeled $(L^s)'$. This decrease in labor supply reflects the substitution effect from the higher marginal tax rate, τ_w.

The tax on labor income does not affect the labor-demand curve, L^d. The reason is that businesses (run by households) still maximize profit by equating the marginal product of labor, MPL, to the real wage rate, *w/P*. For a given *w/P*, the labor-income tax rate, τ_w, does not affect the profit-maximizing choice of labor input, L^d.

We can see from Figure 13.5 that the market-clearing pre-tax real wage rate rises from $(w/P)^*$ to $[(w/P)^*]'$. The market-clearing quantity of labor declines from L^* to $(L^*)'$. We also know that the after-tax real wage rate, $(1-\tau_w)\cdot(w/P)$, must fall overall. That is, the rise in w/P less than fully offsets the decrease in $1-\tau_w$ due to the rise in τ_w. The reason is that L and, hence, the quantity of labor supplied, L^s, have decreased. The quantity of labor supplied will be lower only if $(1-\tau_w)\cdot(w/P)$ has fallen.

We found from our analysis of the labor market that a higher marginal tax rate on labor income, τ_w, lowers the quantity of labor input, L. This effect will spill over to the market for capital services because the reduction in L tends to reduce the schedule for the marginal product of capital services, MPK.

Figure 13.6 shows the effects on the market for capital services. As in Figure 12.3, we plot the real rental price, R/P, on the vertical axis. The reduction in labor input, L, reduces the marginal product of capital services, MPK. The demand for capital services decreases accordingly from the blue curve, labeled $(\kappa K)^d$, to the green one, labeled $[(\kappa K)^d]'$. The supply curve for capital services, $(\kappa K)^s$, shown in red, does not shift. That is, the stock of capital, K, is given, and, for a given real rental price, R/P, suppliers of capital services have no reason to change their utilization rate, κ.

Figure 13.6 shows that the market-clearing real rental price falls from $(R/P)^*$ to $[(R/P)^*]'$. The quantity of capital services falls, because of a decrease in the utilization rate, κ, from $(\kappa K)^*$ to $[(\kappa K)^*]'$. Thus, although the tax rate on labor income, τ_w, does not directly affect the market for capital services, it has an indirect effect on this market. By reducing labor input, L, and thereby decreasing the marginal product of capital services, MPK, an increase in τ_w reduces the quantity of capital services, κK.[8]

Recall that real GDP, Y, is given from the production function as

(13.2) $$Y = A \cdot F(\kappa K, L).$$

In the present case, the technology level, A, does not change. However, we found that a rise in the labor-income tax rate, τ_w, reduced the quantities of labor, L, and capital services, κK. Therefore, Y declines.[9] Thus, our conclusion is that a higher marginal tax rate on labor income, τ_w, leads to a reduction in overall market activity, as gauged by real GDP, Y.

A consumption tax

We show here that a consumption tax at a constant marginal rate, τ_c, has the same effects as the labor-income tax that we studied in the main text. Suppose that real labor income, $(w/P) \cdot L^s$, is untaxed, but an increase in consumption, C, by one unit raises a household's real taxes, T/P, by τ_c units. This tax could be a sales tax or an excise tax or a valued-added tax.

[8] The decrease in κK lowers the schedule for the marginal product of labor, MPL, and leads thereby to a further leftward shift in the labor-demand curve, L^d, in Figure 13.5. This shift leads to a further decrease in L.

[9] We can work out how the decrease in real GDP, Y, divides up between consumption, C, and gross investment, I. We know from chapter 8 that a fall in Y by one unit corresponds to a decrease in real household income by one unit. Since the decrease in income is long-lasting, we predict that the propensity to consume would be close to one. Hence, C would fall by roughly one unit. However, the decrease in R/P implies a fall in the real interest rate, r. This change has an intertemporal-substitution effect, which raises current consumption compared to future consumption. This effect offsets the decrease in C by one unit. Hence, we find that current consumption would decrease overall by less than Y. Since $Y = C+I+G$ and G is fixed, I must decline.

If a household works one unit more, it again gets w/P additional units of real labor income. Our assumption now is that this additional labor income is untaxed. Suppose that the household raises consumption by ΔC units. This change raises consumption taxes by $\tau_c \cdot \Delta C$. Therefore, the extra income of w/P units must cover the added consumption, ΔC, plus the added taxes, $\tau_c \cdot \Delta C$:

$$w/P = \Delta C + \tau_c \cdot \Delta C$$

$$= \Delta C \cdot (1+\tau_c).$$

If we divide through by $1+\tau_c$, we can solve out for the additional consumption:

$$\Delta C = (w/P)/(1+\tau_c). \tag{13.3}$$

Hence, for each unit more of labor—and, therefore, each unit less of leisure—a household gets $(w/P)/(1+\tau_c)$ units more of consumption. For example, if $\tau_c = 0.10$, the household gets about $0.9 \cdot (w/P)$ extra units of consumption. The important point is that the higher τ_c, the worse the deal. Therefore, if τ_c rises, we predict that the household would work less, enjoy more leisure, and consume less.

With a labor-income tax at the marginal rate τ_w, we found that a household's labor supply, L^s, depended on the after-tax real wage rate, $(1-\tau_w) \cdot (w/P)$. With a consumption tax at the marginal rate τ_c, we see from equation (13.3) that the after-tax real wage rate—in terms of the extra consumption that can be bought with an additional unit of labor—is $(w/P)/(1+\tau_c)$. Therefore, L^s depends on $(w/P)/(1+\tau_c)$. Thus, increases in τ_w and τ_c have analogous negative effects on L^s. The conclusion is that consumption taxation has the same economic effects that we found for labor-income taxation.[10]

B. A tax on asset income

Go back to the household budget constraint:

$$(13.1) \qquad C + (1/P)\bullet\Delta B + \Delta K = (w/P)\cdot L^s + r\cdot(B/P + K) + V/P - T/P.$$

Suppose now that taxes depend on a household's asset income, $r\cdot(B/P + K)$. Note that this income equals the real interest payments on bonds, $r\cdot(B/P)$, plus the return on ownership of capital, rK. The term rK equals the net real rental payments on capital, $[(R/P)\cdot\kappa - \delta(\kappa)]\cdot K$, because of the condition that the rates of return on bonds and capital are the same:

$$(13.4) \qquad r = (R/P)\cdot\kappa - \delta(\kappa).$$

Let τ_r be the marginal tax rate on asset income. We are assuming that all forms of asset income are taxed at the same rate. This simplifying assumption will give us the main results about taxation of asset income.[11] Since the tax rate, τ_r, is the same for interest income as for income on ownership of capital, the equality between the two rates of return still holds in equation (13.4).

We know that the real interest rate, r, has an intertemporal-substitution effect on consumption. A reduction in year 1's consumption, C_1, by one unit allows a household to

[10] The results are different if the tax rates vary over time in a predictable way. For example, households are motivated to consume a lot in years in which the consumption tax rate, τ_c, is relatively low. In contrast, anticipated variations in labor-income tax rates, τ_w, affect the time pattern of labor supply.

[11] Real-world tax systems often treat differently various forms of asset income, which include interest, dividends, capital gains, and parts of self-employment income.

raise year 2's consumption, C_2, by 1+r units. Therefore, an increase in r motivates the household to lower C_1 compared to C_2. The difference now is that the additional r units of asset income in year 2 are taxed at the rate τ_r. That is, the added r units of income are offset by an added $\tau_r \cdot r$ units of taxes. Thus, if a household reduces C_1 by one unit, it can raise C_2 by only $1 + (1-\tau_r)\cdot r$ units. What matters, therefore, for the choice between C_1 and C_2 is the **after-tax real interest rate**, $(1-\tau_r)\cdot r$. If τ_r rises, for given r, $(1-\tau_r)\cdot r$ declines. Therefore, households have less incentive to defer consumption, and they react by increasing C_1 compared to C_2. To put it another way, for given real income in year 1, households would consume more and save less.

If we multiply through equation (13.4) by $(1-\tau_r)$, we can relate the after-tax real interest rate to the after-tax return on ownership of capital:

after-tax real interest rate = after-tax rate of return on ownership of capital.

$$(1-\tau_r)\cdot r = (1-\tau_r)\cdot[(R/P)\cdot\kappa - \delta(\kappa)]. \qquad (13.5)$$

Thus, to determine $(1-\tau_r)\cdot r$, we have to know the real rental price, R/P, and the capital utilization rate, κ. These values are determined, as before, by the clearing of the market for capital services.

Go back to Figure 13.6, which considers the demand for and supply of capital services. The first point is that the marginal tax rate on asset income, τ_r, does not affect the demand curve for capital services, $(\kappa K)^d$, which is shown in blue. Since business profit, Π, is not taxed, this curve still comes from the equation of the marginal product of capital services, MPK, to R/P.[12] Therefore, a change in τ_r does not shift $(\kappa K)^d$.

[12] A tax on business profit tends to affect the demand for capital services. For example, in the system of corporate-profits taxation described in n. 5, an increase in the tax rate on corporate profits would reduce this demand.

In chapter 8, we worked out the supply curve for capital services, $(\kappa K)^s$. For a given stock of capital, K, owners of capital chose the utilization rate, κ, to maximize their net rental income:

$$[(R/P)\cdot\kappa - \delta(\kappa)]\cdot K.$$

We assumed that a higher κ resulted in a higher depreciation rate, as represented by the function $\delta(\kappa)$. From this formulation, we found that an increase in the real rental price, R/P, raised the utilization rate, κ, and thereby increased the quantity of capital services supplied, $(\kappa K)^s$. That is why the curve $(\kappa K)^s$, shown in red in Figure 13.6, slopes upward.

With a tax on income from capital at the rate τ_r, owners of capital would seek to maximize their after-tax net rental income, given by

$$(1- \tau_r)\cdot [(R/P)\cdot\kappa - \delta(\kappa)]\cdot K.$$

However, for any τ_r, this maximization is equivalent to the maximization of the term inside the brackets, $[(R/P)\cdot\kappa - \delta(\kappa)]\cdot K$, that is, the same expression as before. Therefore, for a given R/P, the chosen utilization rate, κ, does not depend on the tax rate, τ_r. Since the capital stock, K, is given and since τ_r does not affect κ, we conclude that τ_r does not affect the supply of capital services, $(\kappa K)^s$. Therefore, a change in τ_r does not shift the supply curve, $(\kappa K)^s$, in Figure 13.6.

Since a change in τ_r does not affect the demand and supply curves in Figure 13.6, it does not affect the market-clearing real rental price, $(R/P)^*$, and the quantity of capital services, $(\kappa K)^*$. That is, the capital stock, K, is fixed, and the capital utilization rate, κ, does not change.

Since the quantity of capital services, κK, does not change, there is no effect on the demand curve for labor, L^d, shown in red in Figure 13.5. The supply curve for labor,

L^s, shown in blue, also does not shift. Therefore, the market-clearing real wage rate, $(w/P)^*$, and the quantity of labor, L^*, do not change. Since κK and L are the same, we have from the production function, $Y = A \cdot F(\kappa K, L)$ in equation (13.2), that real GDP, Y, does not change. Thus, our conclusion is that a change in the marginal tax rate on asset income, τ_r, does not affect real GDP. We should stress, however, that this result applies in the short run, when the stock of capital, K, is fixed.

Since the real rental price, R/P, and the capital utilization rate, κ, are unchanged, the pre-tax rate of return on ownership of capital, $(R/P) \cdot \kappa - \delta(\kappa)$, does not change. But then, the after-tax rate of return, $(1-\tau_r) \cdot [(R/P) \cdot \kappa - \delta(\kappa)]$, must fall if τ_r rises.
Equation (13.5) tells us that the after-tax real interest rate, $(1-\tau_r) \cdot r$, equals the after-tax rate of return on ownership of capital:

$$(13.5) \qquad (1-\tau_r) \cdot r = (1-\tau_r) \cdot [(R/P) \cdot \kappa - \delta(\kappa)].$$

Therefore, the increase in τ_r lowers the after-tax real interest rate, $(1-\tau_r) \cdot r$.

The decrease in $(1-\tau_r) \cdot r$ has intertemporal-substitution effects on consumption. Households raise year 1's consumption, C_1, compared to year 2's, C_2. Or, to put it another way, for given real income in year 1, households consume more and save less. Recall, however, that year 1's real GDP, Y_1, does not change. Since $Y_1 = C_1 + I_1 + G_1$ and G_1 is unchanged, the increase in C_1 must correspond to a reduction in year 1's gross investment, I_1. In other words, a higher tax rate, τ_r, on asset income leads to higher C_1 and lower I_1.

In the long run, the reduction in gross investment, I, results in a stock of capital, K, that is smaller than it otherwise would have been. This reduced K will lead to a lower real GDP, Y. Therefore, although an increase in the tax rate on asset income, τ_r, does not affect real GDP in the short run, it decreases real GDP in the long run.[13]

IV. An Increase in Government Purchases Financed by a Labor-Income Tax

In section VI of chapter 12, we studied the economic effects from a permanent increase in government purchases, G. We assumed, unrealistically, that the increase in G was financed by lump-sum taxes. Our finding was that an increase in G by one unit left real GDP, Y, unchanged and reduced consumption, C, by about one unit. Hence, gross investment, I, was unchanged. Also unchanged were the real wage rate, w/P, the real rental price, R/P, and the real interest rate, r.

These results depended on the assumption that labor supply, L^s, was fixed. Now we reconsider this assumption while also allowing additional government purchases, G, to be financed, more realistically, by a tax on real labor income, $(w/P)\cdot L$. In particular, we assume that an increase in the marginal income-tax rate on labor income, τ_w, accompanies a permanent rise in G.

We will get different results from those in chapter 12 if the combination of permanently increased government purchases, G, and a higher marginal income-tax rate, τ_w, affects labor supply, L^s. Therefore, we have to consider the various forces that affect L^s.

[13] It is possible to show that a tax on asset income is equivalent to a time-varying tax on consumption. That is, the asset-income tax implicitly taxes future consumption more heavily than current consumption. In most cases, the economy does better if the government raises a given total of revenue through a consumption tax that is uniform over time, rather than time varying. Thus, the government should prefer a constant tax rate on consumption, τ_c, to a tax on asset income, τ_r.

- We observed in section VI of chapter 12 that an increase by one unit in each year's government purchases, G, requires real taxes less real transfers, $T/P - V/P$, to rise by one unit in each year. Therefore, households have one unit less of real income net of taxes in each year. In response to the negative income effect, households would raise labor supply, L^s, in each year.
- In the box on "useful public services" in section VI of chapter 12, we assumed that government purchases, G, provide public services that yield utility for households. We assumed that each unit of G is equivalent, in terms of utility, to λ units of consumption, C, where λ is greater than zero. When we include the service value of government purchases in households' full income, we get that an increase in G by one unit raises this full income by λ units. The combination of this effect with the rise in real taxes less real transfers, $T/P - V/P$, by one unit implies that households' full income falls by $1-\lambda$ units. If λ is less than one, we still get that full income falls when G rises. Therefore, the negative income effect still predicts that labor supply, L^s, rises each year. However, the higher λ, the weaker is this effect.
- We found in section III.A of this chapter that the substitution effect from a higher marginal tax rate, τ_w, reduces labor supply, L^s. Figure 13.5 shows this effect. Note that this analysis ignored any income effects.

We see that the overall effect from a rise in government purchases, G, on labor supply, L^s, depends on the offsetting influences from an income effect and a substitution effect. The negative income effect predicts that L^s would rise. The substitution effect from the higher tax rate, τ_w, predicts that L^s would fall. The overall effect on L^s is uncertain.

Empirically, the overall effect from permanently increased government purchases, G, on labor supply, L^s, may be small. This interpretation is consistent with the finding in Figure 12.5 that, from 1954.1 to 2004.3, the fluctuations in government purchases were unrelated to the fluctuations in real GDP. We found substantial effects on real GDP only when we looked at the temporary increases in G during major wars, notably World Wars I and II and the Korean War. Thus, in the end, our assumption in section VI of chapter 12 that labor supply was fixed may have been appropriate. In particular, this assumption led to the reasonable conclusion that permanently higher government purchases had a small effect of uncertain sign on real GDP.

The Laffer curve

A permanent increase in government purchases, G, requires a permanent increase in real taxes, T/P. (We assume here that real transfers, V/P, are fixed.) We assumed in the previous section that a rise in T/P went along with a rise in the marginal income-tax rate, τ_w. We can think here of the higher τ_w as taking the form of increases in marginal tax rates at all levels of income in the individual income tax. We explore here the relation between T/P and τ_w.

This relation is called a **Laffer curve**, named after the economist Arthur Laffer.[14]

The real taxes, *T/P*, collected from a tax on labor income can be written as

$$T/P = \left[\frac{(T/P)}{(w/P)\cdot L}\right]\cdot(W/P)\cdot L,$$

real tax revenue = (average tax rate) x (tax base).

The tax base for a labor-income tax is real labor income, $(w/P)\cdot L$. The average tax rate is the ratio of *T/P* to the tax base.

We assume that a higher marginal income-tax rate, τ_w, goes along, in a mechanical way determined by the tax law, with a higher average tax rate. Thus, the overall relation between τ_w and the real taxes collected, *T/P*, depends on the response of the tax base, $(w/P)\cdot L$. In the example that we discussed in the previous section, $(w/P)\cdot L$ did not change when τ_w increased, along with a permanent rise in government purchases, *G*. The idea of the Laffer curve is that the response of $(w/P)\cdot L$ to an increase in τ_w tends to become negative when τ_w is high. Moreover, this response eventually becomes so strong that *T/P* falls when τ_w rises.

When τ_w is zero at all levels of income, the real revenue collected, *T/P*, is also zero. Therefore, the Laffer curve, shown in Figure 13.7, begins at the

[14] For a discussion of the Laffer curve, see Fullerton (1982).

origin. If the government raises τ_w above zero, T/P becomes positive. Therefore, the Laffer curve has a positive slope when τ_w is small.

The negative influence of a higher τ_w on labor, L, works through the substitution effect from the after-tax real wage rate, $(1-\tau_w)\cdot(w/P)$. Think about the term $1-\tau_w$. If $\tau_w = 0$, an increase in τ_w by, say, 0.1 has a relatively small proportionate effect on $1-\tau_w$. This term falls from 1 to 0.9, that is, by 10%. However, if $\tau_w = 0.5$, an increase in τ_w by 0.1 decreases $1-\tau_w$ from 0.5 to 0.4, or by 20%. When $\tau_w = 0.8$, the corresponding effect is 50% (from 0.2 to 0.1), and when $\tau_w = 0.9$, it is 100% (from 0.1 to 0). This arithmetic suggests that labor supply, L^s, would fall eventually as τ_w rose higher and higher. In the extreme, when τ_w rises from 0.9 to 1.0, households may stop working entirely. Eventually, this negative effect on labor supply means that the tax base, $(w/P)\cdot L$, would fall enough to more than offset the rise in the average tax rate. At that point, real tax revenue, T/P, would decline with a further rise in τ_w.

The graph in Figure 13.7 reflects this discussion. The slope of the relation between real tax revenue, T/P, and the marginal tax rate, τ_w, is positive at the origin but becomes flatter as τ_w rises. Eventually, T/P reaches a peak, when τ_w attains the value denoted $(\tau_w)^*$. For still higher marginal tax rates, T/P falls as τ_w rises. The graph assumes that, at the confiscatory tax rate of 100%, real wage income, $(w/P)\cdot L$—at least the part reported to the tax authorities—falls to zero, so that T/P is zero.

In 1980, when Ronald Reagan was elected U.S. president, some advocates of supply-side economics used a picture like the one in Figure 13.7 to argue for an across-the-board cut in U.S. income-tax rates. These economists contended that the average marginal tax rate on income (which we estimated in Figure 13.4) exceeded $(\tau_w)^*$, so that a general cut in tax rates would yield a larger real tax revenue, T/P. However, there is no evidence that the United States has ever reached high enough tax rates for this result to apply.

Lawrence Lindsey (1987) estimated the effect of the Reagan tax cuts from 1982 to 1984 on the tax payments by taxpayers in various income groups. He found that the reductions in tax rates lowered tax collections overall and for taxpayers with middle and low incomes. However, among taxpayers with the highest incomes (adjusted gross incomes above $200,000), the increase of reported taxable incomes was so great that it more than offset the decrease in tax rates. Lindsey estimated that the tax-rate cuts raised collections in this group by 3% in 1982, 9% in 1983, and 23% in 1984. Therefore, although U.S. taxpayers as a whole were not on the falling portion of the Laffer curve in Figure 13.7, the taxpayers with the highest incomes appeared to be in this range.

A study by Charles Stuart (1981) estimated that the maximum of real tax revenue occurred in Sweden at an average marginal tax rate of 70%. That is, he estimated $(\tau_w)^*$ in Figure 13.7 to be about 70%. The actual average marginal tax rate in Sweden reached 70% in the early 1970s and rose subsequently to about 80%.[15] Therefore, Sweden was apparently operating on the falling portion of the Laffer curve during the 1970s. A similar study by A. Van Ravestein and H. Vijlbrief (1988) estimated that $(\tau_w)^*$ was also about 70% in the Netherlands. They found that the actual marginal tax rate in the Netherlands reached 67% in 1985—close to but not quite as high as the estimated $(\tau_w)^*$.

V. Transfer Payments

We assumed up to now that real transfer payments, V/P, were lump sum. But, as with taxes, transfers are not lump sum in the real world. Rather, most transfer programs relate an individual's payments to the individual's characteristics. For example, welfare programs give money or services such as healthcare to poor persons and then reduce or eliminate the transfers if a person's income rises. Similarly, to qualify for unemployment insurance, a recipient must not be working. Until recently, the U.S. social security program reduced the pension payments to an elderly person if that person received labor income above a specified amount. However, the law now allows a person of the normal

[15] These estimates of marginal tax rates include consumption taxes, as well as income taxes.

retirement age or older to earn income without triggering reductions in the person's social security pension.

The general point is that an income-tested transfer program—one that reduces transfers when income rises—effectively imposes a positive marginal income-tax rate on labor income (and possibly also on asset income). Typically, an increase in the scale of a transfer program—for example, an expansion of the welfare system—leads to an increase in the average marginal tax rate, τ_w, implied by the program. Therefore, if we want to analyze the economic effects from increased real transfers, V/P, we should take account of this increase in τ_w.

Suppose that the government increases real transfers, V/P, and finances these expenditures with increased real taxes, T/P, collected by the individual income tax. In this case, the average marginal income-tax rate, τ_w, rises for two reasons. First, as we noted in section III.A, a rise in T/P typically goes along with a higher marginal-income rate on persons who are paying individual income taxes. Second, for people who are receiving transfers—such as poor welfare recipients—the expansion of the transfer program tends to raise the implicit marginal income-tax rate, τ_w. That is, these persons tend to lose a larger amount of transfers when they earn more income. Overall, we get an increase in τ_w from both sides. Thus, we predict even stronger adverse economic consequences of the type analyzed in section III.A. In particular, labor input, L, capital services, κK, and real GDP, Y, tend to decline. The effects are stronger than those found before because we are now including the additional effect on the average marginal income-tax rate from the incentives contained in the transfer programs.

VI. Summing Up

In the previous chapter, the government used lump-sum taxes. This chapter considers realistic forms of taxes, which affect incentives to work and save. A tax on labor income has a substitution effect that discourages labor supply. Consequently, an increase in the marginal tax rate on labor income reduces the after-tax real wage rate and the quantity of labor. Therefore, real GDP declines. An increase in the marginal tax rate on asset income does not affect real GDP in the short run. However, the decrease in the after-tax real interest rate causes a shift toward consumption and away from investment. Consequently, in the long run, the capital stock and real GDP decline.

We reconsidered a permanent increase in government purchases to include an accompanying rise in the marginal tax rate on labor income. The overall effect on labor supply is ambiguous—the negative income effect suggests higher labor supply, but the substitution effect from the higher tax rate suggests lower labor supply. The evidence is consistent with little net change in labor supply. For example, the stability of labor supply is consistent with the finding that peacetime variations in government purchases play a minor role in economic fluctuations.

Questions and Problems

Mainly for review

13.1. Distinguish between the average tax rate and the marginal tax rate. Must the two be equal for a flat-rate tax?

13.2. Could an increase in the tax rate on labor income reduce real tax revenue? How does the answer depend on the responsiveness of labor supply to the after-tax real wage rate?

13.3. What is the income effect from a rise in the tax rate on labor income, τ_w? Why did we assume no income effect in part of our analysis?

Problems for discussion

13.x. The flat-rate tax

Some U.S. economists advocate shifting from the graduated individual income tax to a flat-rate tax. Under the new system, there would be few deductions from taxable income, and the marginal tax rate would be constant. Because of the elimination of the deductions, the average marginal tax rate would be lower than that under the current system.

a. What would be the economics effects from a shift to a flat-rate tax on labor income?

b. How does the proposed flat-rate tax compare with the present social security payroll tax?

13.x. A consumption tax

Suppose that consumption is taxed in each year at the constant rate τ_c.

a. What is the household budget constraint?

b. What are the effects of an increase in τ_c on the labor market? How do the results compare with those shown for an increase in the tax rate on labor income, τ_w, in Figure 13.5?

c. What are the effects of an increase in τ_c on the market for capital services? How do the results compare with those shown for an increase in the tax rate on labor income, τ_w, in Figure 13.6?

d. Suppose now that τ_c rises in year 1 but does not change in future years. How does this change affect the choice of consumption over time? How would the effects resemble those from a change in the tax rate on asset income, τ_r? In what ways do the effects differ from those from a change in the tax rate on asset income, τ_r?

13.x. Effects of inflation on a graduate-rate income tax

Back in 1985, before the tax simplification of 1986, the individual income-tax system in the United States had many different tax-rate brackets. A married couple paid individual income tax on labor income in accordance with the following table:

Range of taxable income ($)	**Marginal income-tax rate (%)**
3,540-5,719	11

5,720-7,919	12
7,920-12,389	13
12,390-16,649	16
16,650-21,019	18
21,020-25,599	22
25,600-31,119	25
31,120-36,629	28
36,630-47,669	33
47,670-62,449	38
62,450-89,089	42
89,090-113,859	45
113,860-169,019	49
169,020-	50

a. Suppose that each person's real income stays constant over time, so that inflation steadily raises each person's nominal income. If the tax schedule shown in the table had remained unchanged, what would have happened over time to each couple's marginal income-tax rate?

b. Assume now that the dollar bracket limits shown in the left column of the table are adjusted proportionately (or "indexed") over time for changes in the price level. That is, if the price level rises by 5%, each dollar amount rises by 5%. What then is the effect of inflation on each couple's marginal income-tax rate? (This indexing provision applies in the United States since 1985.)

13.x. Inflation and taxes on asset income

Suppose that the tax rate on asset income is τ_r. Suppose (as is true in the United States) that the tax is levied on nominal interest income. Assume, however (as it not entirely accurate), that the tax is levied on real returns on capital.

a. What is the after-tax real interest rate on bonds?

Consider a permanent, unanticipated increase in the money growth rate from μ to μ'. Assume that τ_r does not change.

b. What is the effect on the inflation rate, π?

c. What is the effect on the after-tax real return on capital?

d. What is the effect on the after-tax real interest rate on bonds? What happens to the nominal interest rate, i? Does it move one-for-one with π?

13.x. Effects of transfer programs on labor supply

Discuss the effects on labor supply from the following governmental programs:

a. The food stamp program, which provides subsidized coupons for purchases of food. The allowable subsidies vary inversely with family income.

b. A negative income tax, such as the earned-income tax credit (EITC). This program provides cash transfers to poor persons. In the EITC, the amount transferred initially rises with labor income. Then, subsequently, the amount paid falls with labor income.

c. Unemployment insurance, which gives cash payments over a specified interval to persons who have lost jobs and are looking for work. In the United States, the benefits typically last as long as six months; many European countries have longer periods of eligibility.

d. Retirement benefits paid under social security. What is the consequence of the income-test provision? Before 1972, this test specified that an increase by two dollars in labor income causes a reduction by one dollar in retirement benefits. Since 2000, the test applies only to retirees below normal retirement age (currently ages 66-67) and has two ranges, one where two dollars of income reduce benefits by one dollar and a second where three dollars of income reduce benefits by one dollar.

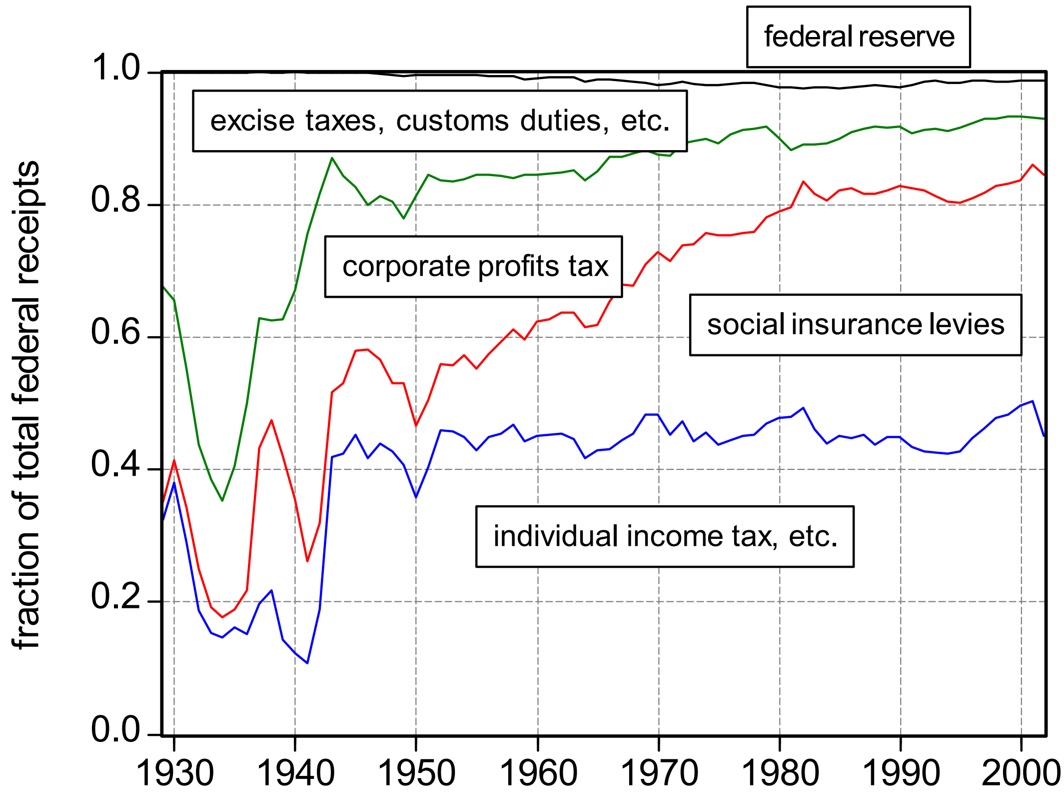


Figure 13.1

Breakdown of Federal Government Receipts

The figure shows the fractions of federal government receipts in five categories. These are individual income taxes, etc. (including estate taxes and personal non-tax payments); contributions for social insurance (which include payments by employers and employees into pension funds and Medicare, and unemployment-insurance premiums); corporate profits taxes; excise taxes, customs duties, etc. (including taxes on petroleum products, alcohol, and tobacco, and non-tax payments, such as deposit-insurance premiums); and the Federal Reserve's transfers to the U.S. Treasury (which are treated in the national accounts as a part of corporate profits taxes).

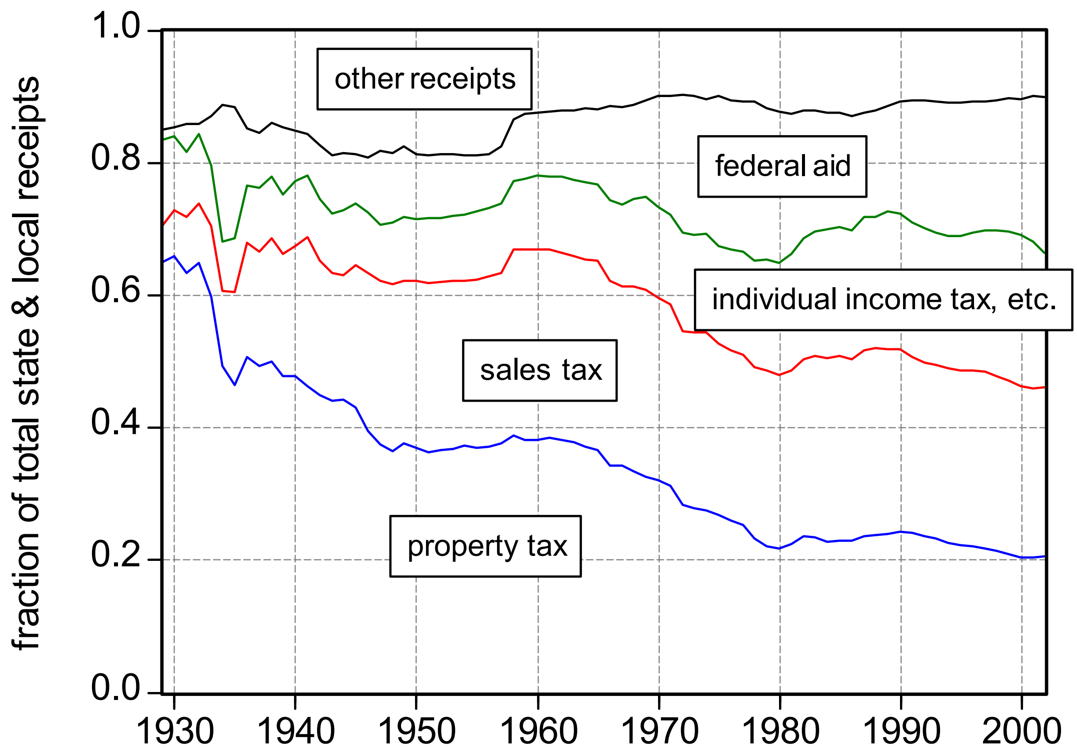


Figure 13.2

Breakdown of State & Local Government Receipts

The figure shows the fractions of state & local government receipts in five categories. These are property taxes; sales taxes; individual income taxes, etc. (including estate taxes and personal non-tax payments); federal aid (transfers from the federal government to state & local governments); and other receipts (corporate profits taxes, social insurance contributions, and miscellaneous fees and taxes).

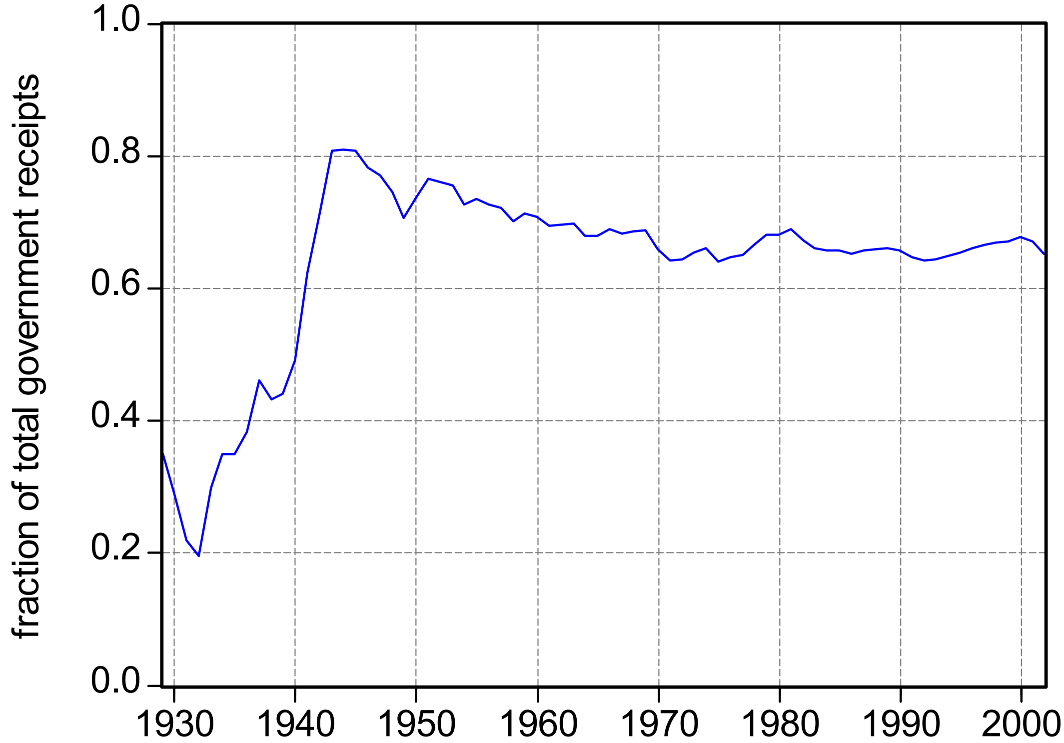


Figure 13.3

Federal Receipts as a Fraction of Total Government Receipts

The graph shows the ratio of federal government receipts to total government receipts. Total receipts are federal receipts plus state & local receipts less transfers from the federal government to state & local governments.

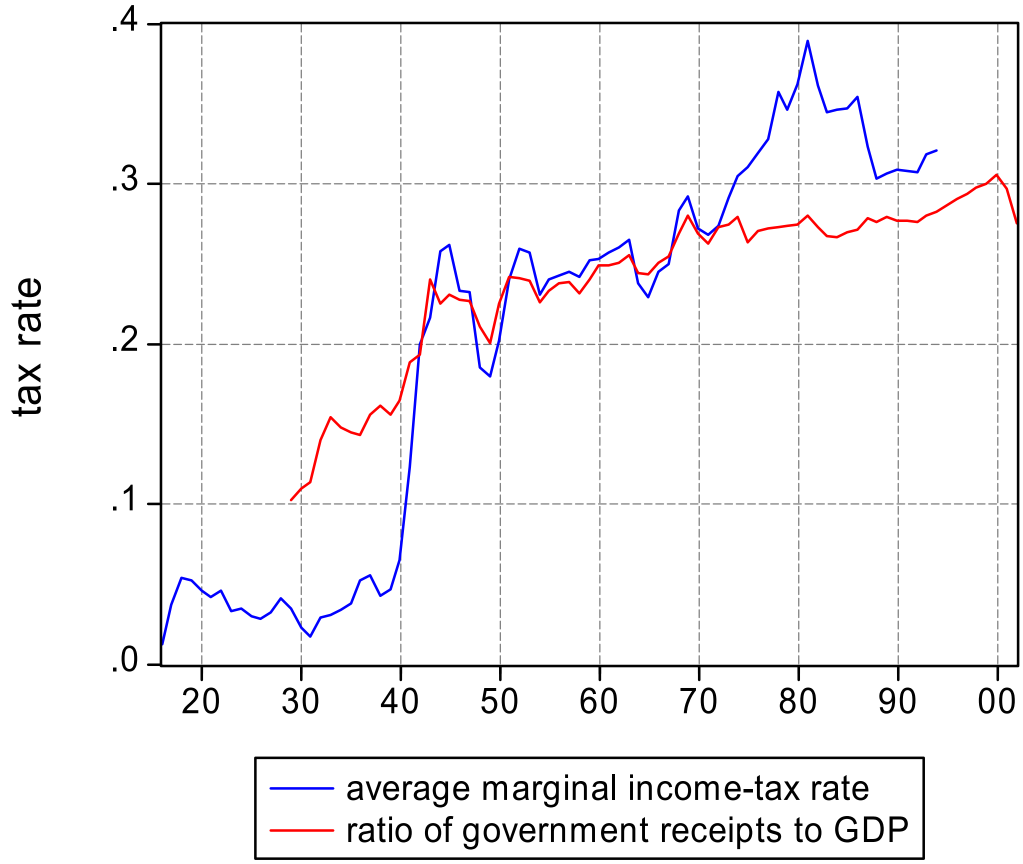


Figure 13.4

Measures of Tax Rates

The red graph is the ratio to GDP of total government receipts. Total receipts are federal receipts plus state & local receipts less transfers from the federal government to state & local governments. This ratio would measure the average tax rate if all taxes were levied on GDP. The blue graph is the income-weighted average marginal tax rate from the federal individual income tax and social security. This variable was estimated in Barro and Sahasakul (1983, 1986) and was updated to 1994 by Mulligan (2003).

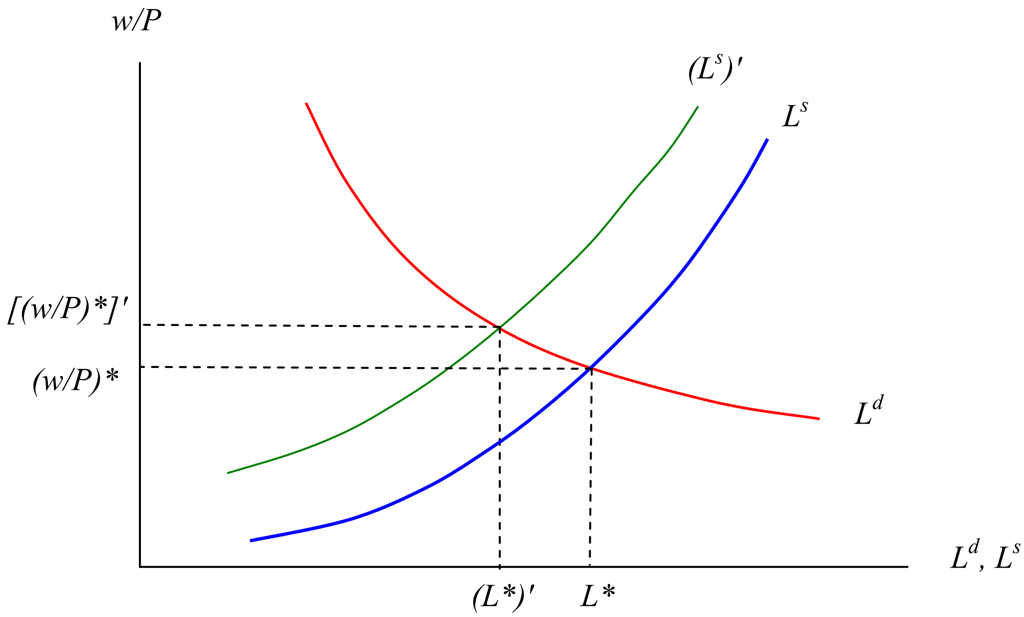


Figure 13.5

Effect of an Increase in the Labor-Income Tax Rate on the Labor Market

The downward-sloping labor-demand curve, L^d, shown in red, comes from Figure 12.6. The upward-sloping labor-supply curve, L^s, shown in blue, also comes from Figure 12.6. An increase in the marginal tax rate on labor income, τ_w, shifts the labor-supply curve leftward to the green one, $(L^s)'$. Consequently, the market-clearing before-tax real wage rate rises from $(w/P)^*$ to $[(w/P)^*]'$. The market-clearing quantity of labor input falls from L^* to $(L^*)'$.

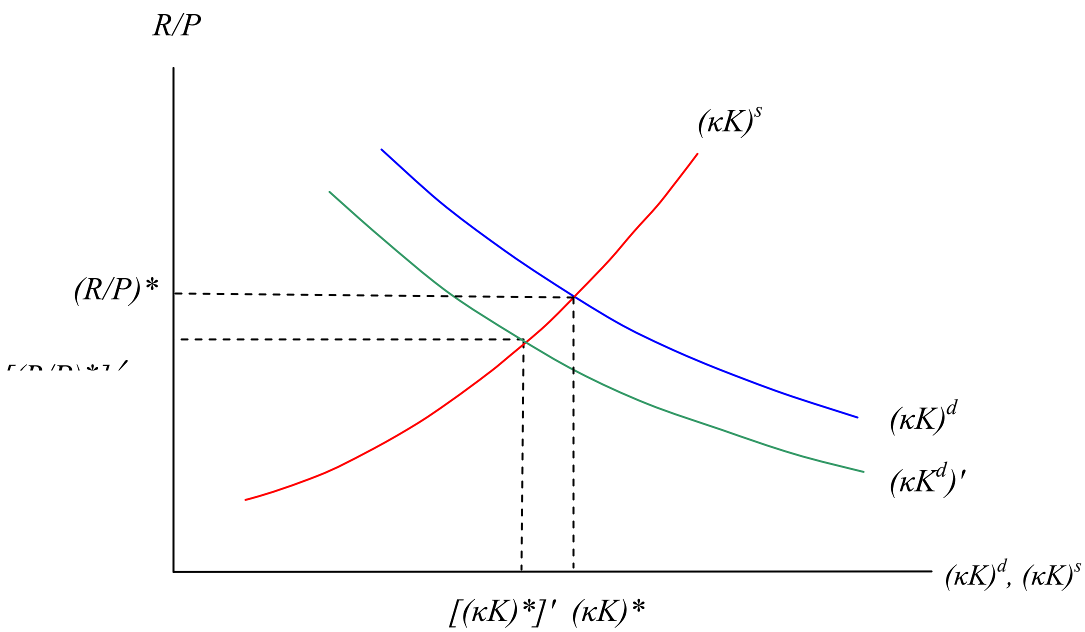


Figure 13.6

Effect of an Increase in the Labor-Income Tax Rate on the Market for Capital Services

This construction comes from Figure 12.3. The reduction in employment from L^* to $(L^*)'$, shown in Figure 13.5, reduces the schedule for the marginal product of capital services, MPK. Therefore, the demand curve for capital services shifts leftward, from $(\kappa K)^d$, shown in blue, to $[(\kappa K)^d]'$, shown in green. The supply curve for capital services, $(\kappa K)^s$, shown in red, does not shift. Consequently, the market-clearing real rental price of capital falls from $(R/P)^*$ to $[(R/P)^*]'$. The quantity of capital services declines from $(\kappa K)^*$ to $[(\kappa K)^*]'$. This decrease in capital services corresponds, for a given capital stock, K, to a reduction in the capital utilization rate, κ.

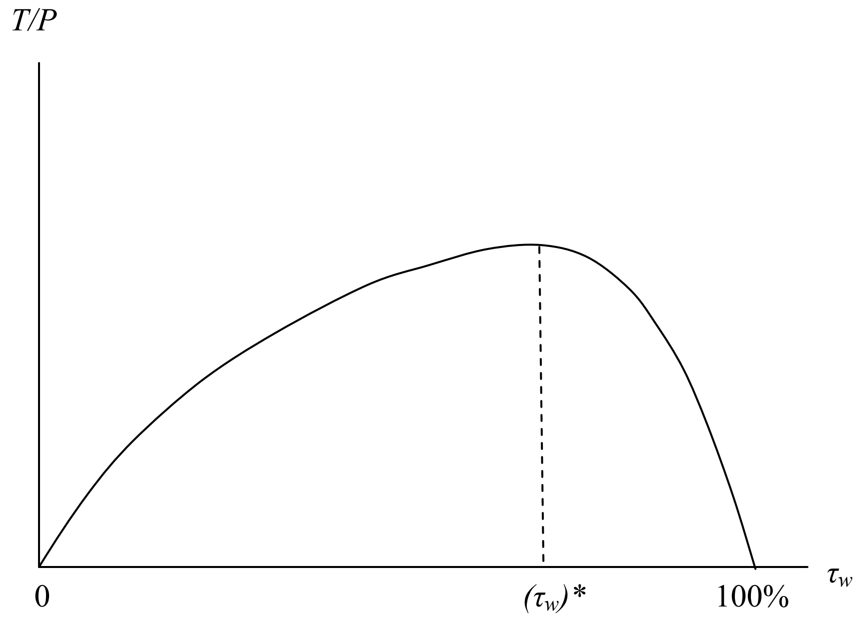


Figure 13.7

The Relation between Tax Receipts and the Marginal Income-Tax Rate (a Laffer Curve)

The horizontal axis has the marginal tax rate on labor income, τ_w. The vertical axis has real tax receipts, T/P. Starting from zero, an increase in τ_w raises T/P. However, as τ_w rises, the slope gets less steep. Eventually, T/P reaches its peak when τ_w reaches $(\tau_w)^*$. Beyond that point, T/P falls as τ_w rises toward 100%.

January 12, 2005

Chapter 14

The Public Debt

In recent years, one of the most controversial economic issues has been the **government's budget deficit**. At least from reading the newspapers, we would think that the economy suffers greatly when the government runs a deficit. The most important task in this chapter will be to evaluate this view. As we shall see, our conclusions depart dramatically from those expressed in the newspapers.

Budget deficits arise when the government finances part of its expenditures by issuing interest-bearing government bonds—**public debt**—rather than levying taxes. A budget deficit means that the quantity of public debt increases over time.

We consider first the history of the public debt in the United States and the United Kingdom. With these facts as a background, we extend the model to allow for public debt. Now the government can run budget deficits rather than levy taxes. We use the model to assess the effects of budget deficits and public debt on economic variables, including real GDP, saving, investment, and interest rates.

I. The History of the Public Debt in the United States and the United Kingdom

Table 14.1 shows the long-term history of the public debt in the United States and the United Kingdom. The table shows the nominal quantity of interest-bearing debt, denoted by B^g, and the ratio of this debt to nominal GDP (or GNP in earlier years).[1] Figure 14.1 shows the debt-GDP ratio for the United States from 1790 to 2002, and Figure 14.2 shows the ratio for the United Kingdom from 1700 to 2001.

For the United States, the major peaks in the ratio of public debt to GDP reflected the financing of wartime expenditures: the ratio reached 0.33 in 1790 after the Revolutionary War, 0.13 in 1816 after the War of 1812, 0.28 in 1866 after the Civil War, 0.31 in 1919 after World War I, and 1.09 in 1946 after World War II. Other wars—such as the Spanish-American War, the Korean War, the Vietnam War, and the Iraq War—were too small to have major effects on the debt-GDP ratio.

Another major influence on the debt-GDP ratio is the business cycle. In recessions, the debt-GDP ratio typically rises, partly because of a drop in GDP and partly because of an increase in debt. A dramatic example is the Great Depression, during which the debt-GDP ratio increased from 0.14 in 1929 to 0.38 in 1933. Qualitatively similar behavior applies to less severe contractions. For example, the ratio rose from 0.25 in 1979 to 0.32 in 1983 and from 0.042 in 1892 to 0.060 in 1896.

During peacetime, non-recession years, the ratio of public debt to GDP typically declined. For example, after World War II, the ratio fell from 1.09 in 1946 to 0.23 in 1974. The ratio also fell after World War I: from 0.31 in 1919 to 0.14 in 1929, after the Civil War: from 0.28 in 1866 to 0.018 in 1916, after the War of 1812: from 0.13 in 1816 to 0.015 in 1859, and after the Revolutionary War: from 0.33 in 1790 to 0.075 in 1811.

[1] The U.S. figures are net of holdings of public debt by U.S. government agencies and trust funds and the Federal Reserve. For example, in June 2004, of the $7.3 trillion of public debt securities, $2.3 trillion was held by agencies and trust funds and $0.7 trillion by the Federal Reserve. For the United Kingdom, the debt figures are the gross sterling debt of the central government. Net numbers are unavailable for the long-term history. In March 2003, the gross sterling debt of £391 billion corresponded to a public sector net debt of £333 billion.

Particularly interesting is the running down of the debt to near zero from 1834 to 1836. Also noteworthy is the period from 1983 to 1993, when the debt-GDP ratio rose from 0.32 to 0.49 despite the absence of war or substantial recession. However, the ratio declined to 0.33 in 2001, before rising along with the recession to 0.34 in 2002. This recent ratio is not high by historical standards.

The experience of the United Kingdom is broadly similar to that for the United States. The major peaks in the debt-GDP ratio were again associated with wartime: 0.50 after the Wars of Spanish and Austrian Succession in 1722, 1.1 at the end of the Seven Years' War in 1764, 1.2 after the War of American Independence in 1785, 1.3 after the Napoleonic Wars in 1816, 1.4 at the end of World War I in 1919, and 2.6 after World War II in 1946. It is noteworthy that the high points for the U.K. debt in relation to GDP were more than twice as great as those for the United States. It is also interesting that the public debt constituted more than 100% of annual GDP as long ago as the 1760s. Large amounts of public debt are not a modern invention![2]

Recessions again had a positive impact on the debt-GDP ratio. This response shows up especially during the economically depressed periods from 1920 to 1923 and 1929 to 1933. The debt-GDP ratio rose from 1.2 to 1.7 during the first interval and from 1.5 to 1.8 during the second.

Periods with neither war nor recession typically displayed declining debt-GDP ratios. For example, the ratio declined from 2.57 in 1946 to 0.35 in 1990. The most recent observation, 0.37 in 2001, is similar to that in 1990.

[2] Some researchers argue that Britain learned how to issue these large amounts of public debt with the Glorious Revolution of 1688. See North and Weingast (1989) and Sargent and Velde (1995). This method of financing may have given Britain a great advantage over France in the fighting of numerous wars.

II. Characteristics of Government Bonds

In the model, the government can now borrow funds from households by issuing interest-bearing bonds. We assume that these government bonds pay interest and principal in the same way as the private bonds in the model. In particular, we continue to simplify by assuming that all bonds have very short maturity.[3] We also assume that, like private bonds, government bonds specify nominal amounts of principal and interest. That is, we do not consider indexed bonds, which we discussed in chapter 11.[4]

We assume that bondholders (households in our model) regard government bonds as equivalent to private bonds. Specifically, we do not allow for the possibility that private bonds are riskier than government bonds in terms of the probability of default. In this case, households would hold the two kinds of bonds only if they paid the same nominal interest rate, i. Therefore, we continue to assume in the model that there is only one nominal interest rate, i, paid on all bonds.

Our assumption about public and private bonds contrasts with our treatment of money. Because currency pays no nominal interest, households (or businesses) would likely find it profitable to issue money. In particular, the higher the nominal interest rate, i, the greater the gain from entering the business of creating currency. However, because of legal restrictions or technical advantages for the government in providing a medium of exchange, we assume that households do not issue currency.

---

[3] In the United States, the average maturity of marketable, interest-bearing U.S. government bonds held by the public was around 9 years in 1946. This maturity declined to a low point of 2-1/2 years in 1976. During most of this period, the U.S. Treasury was prohibited from issuing long-term bonds at interest rates that would have made them marketable. With the ending of this restriction, the average maturity rose to about six years in 1990. In 2004, the maturity was around five years. For the United Kingdom, in 2003, 36% of marketable government bonds had a maturity up to 5 years, 35% was between 5 and 15 years, and 29% was greater than 15 years.

[4] In 2004, indexed bonds constituted about 5% of the total of U.S. government bonds held by the public. For the United Kingdom, in 2003, indexed bonds were much more important—19% of the outstanding gross sterling debt of the central government.

With respect to interest-bearing bonds, we assume no legal restrictions on private issues and no technical advantages for the government. Hence, we do not allow the interest rate on government bonds to differ from that on private bonds. This assumption accords reasonably well with the U.S. data if we interpret private bonds as prime corporate obligations. For example, the market yield on three-month maturity U.S. Treasury Bills averaged 4.8% between 1947 and 2003, while that on three-month maturity commercial paper (short-term debt issued by the most credit-worthy businesses) averaged 5.5%.[5] A similar comparison applies for long-term bonds.

Denote by B_t^g the nominal amount of government bonds outstanding at the end of year *t*. We still use the symbol B_t for private bonds (issued by households in our model). Thus, a household's total holding of bonds at the end of year *t* is

$$\textit{total bond holdings} = \textit{private bonds} + \textit{government bonds}$$

$$= B_t + B_t^g .$$

The quantity of private bonds held by all households is still zero, because the positive amount held by one household must correspond to the debt of another household. Therefore, $B_t = 0$ still holds in the aggregate. This result means that the total quantity of bonds held by all households equals the public debt, B_t^g:

$$\textit{total bond holdings of all households} = B_t^g .$$

[5] Some of the difference in yields reflects two advantages of the government's bonds: first, the interest payments are exempt from state and local income taxes, and, second, holdings of U.S. Treasury Bills satisfy the requirement that commercial banks hold government bonds as backing for the government's deposits in these banks.

We usually think of situations in which the government is a net debtor to the private sector, so that B_t^g is greater than zero. In principle, however, the government could be a creditor, in which case B_t^g would be less than zero.[6]

III. Budget Constraints and Budget Deficits

To consider budget deficits and public debt, we have to see how they fit into the government's budget constraint. We begin by extending the scope of the government's budget constraint.

A. The government's budget constraint

We introduced the government's budget constraint for year *t* in chapter 12:

$$G_t + V_t/P_t = T_t/P_t + (M_t - M_{t-1})/P_t. \tag{12.1}$$

government purchases + transfer payments = taxes + money issue.

Recall that all of the items are in real terms, measured as goods per year.

The presence of public debt alters the government's budget constraint in two ways. First, we have to include the interest payments on the outstanding stock of government bonds. This amount is $i_{t-1} \cdot B_{t-1}^g$ in nominal units. The real value of these interest payments, $i_{t-1} \cdot (B_{t-1}^g/P_t)$, adds to the government's uses of funds on the left-hand side of equation (12.1).

[6] The last time this became a serious possibility for the United States was in 1834-36, when the public debt came close to zero—see Figure 14.1. A major concern at the time was the outlet for future government surpluses once the national debt was fully paid off—see the discussion in Dewey (1931, p. 221). Alan Greenspan, the head of the U.S. Federal Reserve, expressed concern in 2001 that the United States would return to this situation because of large projected budget surpluses. However, the return of budget deficits shortly thereafter made this issue academic.

The second new consideration is that the net issue of debt during year t is a source of funds for the government. The dollar amount of this debt issue is $B_t^g - B_{t-1}^g$.[7] The real value, $(B_t^g - B_{t-1}^g)/P_t$, adds to the government's sources of funds on the right-hand side of equation (12.1). Notice that the term $(B_t^g - B_{t-1}^g)/P_t$ looks like the real revenue from printing money, $(M_t-M_{t-1})/P_t$. In this respect, the printing of money and the printing of interest-bearing bonds play the same role in the financing of the government's expenditures.

When we introduce the two new terms into equation (12.1), we get an expanded form of the government's budget constraint:

Key equation (expanded government budget constraint):

purchases + transfers + interest payments = taxes + debt issue + money issue

$$G_t + V_t/P_t + i_{t-1}\cdot(B_{t-1}^g/P_t) = T_t/P_t + (B_t^g - B_{t-1}^g)/P_t + (M_t-M_{t-1})/P_t. \qquad (14.1)$$

The two new items are the interest payments, $i_{t-1}\cdot(B_{t-1}^g/P_t)$, on the left-hand side and debt issue, $(B_t^g - B_{t-1}^g)/P_t$, on the right-hand side.

To begin our discussion of budget deficits, we return to the simplifying assumptions made in chapters 12 and 13. We assume, first, that the quantity of money, M_t, equals a constant, M. Therefore, the revenue from money creation, $(M_t - M_{t-1})/P_t$, is zero. Second, we neglect inflation—hence, the price level, P_t, equals the constant P. In

[7] Notice that the rolling over or reissue of bonds as they come due is not a net source of funds for the government. What matters is the difference between the stock outstanding at the end of the year, B_t^g, compared to the amount outstanding at the end of the previous year, B_{t-1}^g.

this case, the nominal interest rate, i_t, equals the real interest rate, r_t. Using these assumptions, the government's budget constraint simplifies from equation (14.1) to

$$(14.2) \qquad G_t + V_t/P + r_{t-1}\cdot(B_{t-1}^g/P) = T_t/P + (B_t^g - B_{t-1}^g)/P.$$

Notice that, starting from equation (14.1), we replaced P_t by P and i_{t-1} by r_{t-1}, and we eliminated the term $(M_t-M_{t-1})/P_t$.

B. The budget deficit

To define and calculate the government's budget deficit, it is best to start by thinking about how much the government saves or dissaves. We can think of real saving for the government in the same way as we did for households in chapter 6. If the government saves, its net real assets rise, and if the government dissaves, its net real assets fall. To use these ideas, we have to define the government's net real assets.

The real public debt, B_t^g/P, is a liability of the government. When B_t^g/P rises, the government owes more, therefore has more liabilities and fewer net real assets. Therefore, on this ground, an increase in the real public debt, $(B_t^g - B_{t-1}^g)/P$, means that the government is saving less or dissaving more in real terms.

The real quantity of money outstanding, M_t/P, is also a liability of the government. Thus, an increase in M_t/P means that the government has more liabilities and fewer net real assets. Consequently, on this ground, an increase in real money balances, $(M_t-M_{t-1})/P$, means that the government is saving less or dissaving more in real terms. However, since we have assumed that M_t equals the constant M, the term $(M_t-M_{t-1})/P$ equals zero. Therefore, we can ignore this term.

If the government owned capital, its net real assets would include this capital. In that case, an increase in government owned capital stock—called **net public investment**—would mean that the government had more net real assets. Thus, an increase in net public investment means that the government is saving more or dissaving less in real terms. However, in our model, the government does not own capital. Therefore, net public investment is zero, and we do not have to include this investment as part of the government's saving.

Since the money stock is constant and the government owns no capital, the government's real saving or dissaving equals the negative of the change in the real public debt. If the real public debt increases, the government's real saving is less than zero, and the government is dissaving. If the real public debt decreases, the government's real saving is greater than zero. In terms of an equation, we have

$$\text{(14.3)} \qquad \textit{real government saving} = -(B_t^g - B_{t-1}^g)/P.$$

We can rearrange equation (14.2) to relate real government saving, $-(B_t^g - B_{t-1}^g)/P$, to the government's real expenditure and tax revenue:

$$\text{(14.4)} \qquad -(B_t^g - B_{t-1}^g)/P = T_t/P - G_t - V_t/P - r_{t-1}\cdot(B_{t-1}^g/P).$$

real government saving = real tax revenue – real government expenditure,

where real government expenditure is the sum of purchases, G_t, transfers, V_t/P, and interest payments, $r_{t-1}\cdot(B_{t-1}^g/P)$. When real tax revenue is greater than real government expenditure, real government saving is greater than zero, and the real public debt falls over time.

If the right-hand side of equation (14.4) is greater than zero, the government's revenue exceeds its expenditure. Economists describe this situation as a **budget surplus.**

Thus, the real surplus is the same as the government's real saving. Conversely, if the right-hand side is less than zero, the government's expenditure exceeds its revenue. Economists call this a **budget deficit**. The real deficit is the same as the government's real dissaving. If the right-hand side of equation (14.4) equals zero, the government has a **balanced budget**. In this case, the government's real saving equals zero.

Inflation and the budget deficit in the national accounts

We have assumed that the inflation rate is zero, so that the price level, P_t, equals the constant P. In this case, the national-accounts measure of the government's real saving corresponds to equation (14.4), and the negative of this real saving equals the real budget deficit. However, the national-accounts measure is incorrect when the inflation rate is not zero. To see why, we allow the price level, P_t, to vary over time. To simplify, we still neglect the small part of government revenue that comes from the printing of money. That is, we still assume that M_t equals the constant M.

In the national accounts, nominal government saving equals nominal government revenue less nominal government expenditure:

(14.5) $$nominal\ government\ saving\ (national\ accounts) = T_t - P_tG_t - V_t - i_{t-1} \cdot B_{t-1}^g .$$

Notice that the nominal interest payments, $i_{t-1} \cdot B_{t-1}^g$, are calculated by using the nominal interest rate, i_{t-1}, rather than the real rate, r_{t-1}.

If nominal government revenue, T_t, exceeds nominal expenditure, $P_tG_t + V_t + i_{t-1}\cdot B^g_{t-1}$, the government's nominal debt, B^g, falls over time. Conversely, if T_t is less than $P_tG_t + V_t + i_{t-1}\cdot B^g_{t-1}$, B^g rises over time. That is, the government's budget constraint in nominal terms is

$$(14.6) \qquad -(B^g_t - B^g_{t-1}) = T_t - P_tG_t - V_t - i_{t-1}\cdot B^g_{t-1}.$$

If we divide both sides of the equation by P_t, we get

$$(14.7) \qquad -(B^g_t - B^g_{t-1})/P_t = T_t/P_t - G_t - V_t/P_t - i_{t-1}\cdot(B^g_{t-1}/P_t).$$

The right-hand side is the national-accounts version of the real budget surplus. The problem is that the left-hand side does not equal the change in the government's net real assets—that is, the negative of the change in the real public debt—unless P_t is constant.

The change in the government's net real assets equals the negative of the change in the real public debt:

$$\textit{change in government's net real assets} = -(B^g_t/P_t - B^g_{t-1}/P_{t-1}).$$

If we take $(1/P_t)$ outside of this expression, we get

$$\textit{change in government's net real assets} = -(1/P_t)\cdot[B^g_t - B^g_{t-1}\cdot(P_t/P_{t-1})].$$

Recall that the inflation rate is

$$\pi_{t-1} = (P_t - P_{t-1})/P_{t-1}$$

$$= P_t/P_{t-1} - 1.$$

Therefore, if we rearrange terms, the ratio P_t/P_{t-1} is given by

$$P_t/P_{t-1} = 1 + \pi_{t-1}.$$

If we substitute $P_t/P_{t-1} = 1 + \pi_{t-1}$ into the formula for the change in the government's net real assets, we get

$$\textit{change in government's net real assets} = -(1/P_t)\cdot[B_t^g - B_{t-1}^g\cdot(1+\pi_{t-1})].$$

If we multiply out the terms in B_{t-1}^g, we get

$$\textit{change in government's net real assets} = -(1/P_t)\cdot(B_t^g - B_{t-1}^g - \pi_{t-1}\cdot B_{t-1}^g)$$

$$= -(1/P_t)\cdot(B_t^g - B_{t-1}^g) + \pi_{t-1}\cdot(B_{t-1}^g/P_t).$$

Thus, if we go back to equation (14.7), we find that we have to add the term $\pi_{t-1}\cdot(B_{t-1}^g/P_t)$ to the left-hand side to get the change in the government's net real assets. The reason is that, if π_{t-1} is greater than zero, the rise in the price level from P_{t-1} to P_t reduces the real value of the previous year's nominal debt, B_{t-1}^g, and thereby adds to the government's net real assets.

If we add $\pi_{t-1}\cdot(B_{t-1}^g/P_t)$ to both sides of equation (14.7), the left-hand side equals the change in the government's net real assets. Therefore, we have

$$\textit{change in government's net real assets} = T_t/P_t - G_t - V_t/P_t - i_{t-1}\cdot(B_{t-1}^g/P_t) + \pi_{t-1}\cdot(B_{t-1}^g/P_t).$$

If we combine the two terms in B_{t-1}^g/P_t on the right and use the condition $i_{t-1} - \pi_{t-1} = r_{t-1}$, we get

$$\textit{change in government's net real assets} = T_t/P_t - G_t - V_t/P_t - (r_{t-1})\cdot(B_{t-1}^g/P_t). \tag{14.8}$$

The right-hand side is the same as that in equation (14.4). Therefore, even when π_{t-1} is nonzero, we get the right answer for the change in the government's net real assets if we calculate the government's interest payments with the real interest rate, r_{t-1}, rather than the nominal rate, i_{t-1}. The mistake in the national accounts stems from the use of i_{t-1} to calculate the government's interest payments. This error is important when the inflation rate is high, for example, from the late 1960s through the early 1980s. However, the mistake is less important at times of low inflation, such as the mid 1980s through the early 2000s.

Figure 14.3 shows the history of real budget deficits for the U.S. federal government from 1954 to 2004. We first compute the real public debt for each year as the ratio of the nominal debt (the concept shown in the first column of Table 14.1) to the GDP deflator. Then we calculate the real budget deficit as the change during each year of the real public debt. The preceding box discusses why this concept of the budget deficit is preferable to the standard national-accounts measure when the inflation rate is not zero. To see how the real budget deficit compares with the size of the economy, we divide the real deficit by real GDP.

Figure 14.3 shows that the highest ratios of the real budget deficit to real GDP were those around 4% in 1983-86 and again in 1992. These deficits correspond in Figure 14.1 to the rising debt-GDP ratios associated with the "Reagan-Bush deficits." The rest of the 1990s showed a steady movement toward budget surplus, culminating in the peak ratio above 3% in magnitude in 2000. Then the federal government moved back to a real budget deficit, reaching almost 3% of real GDP in 2003-04.

C. Public saving, private saving, and national saving

To assess the economic effects of budget deficits, it is useful to organize our thinking in terms of effects on saving. However, we have to keep straight three concepts of saving: government (or public) saving, household (or private) saving, and national (or total) saving.

To begin, real government saving is given by

$$(14.3) \qquad \textit{real government saving} = -(B_t^g - B_{t-1}^g)/P.$$

Therefore, if the government runs a real budget deficit, so that $(B_t^g - B_{t-1}^g)/P$ is greater than zero, real government saving is less than zero. We can also note that, if the government owned capital, the change in this capital stock would add to government saving. That is, real government saving would be the change in publicly owned capital—or public net investment—less the government's real budget deficit.

We know from chapter 6 that real household saving equals the change in real household assets. In our previous analysis, these real assets consisted of private bonds, B_t/P; money, M_t/P; and capital, K_t. Now we have to add real assets held as government bonds, B_t^g/P. The economy-wide total of private bonds, B_t, still equals zero. Therefore, if we add up over all households, the change in these bonds also equals zero. If we continue to assume that money, M_t, equals a constant, M, the change in M_t/P when we add up over all households also equals zero. Therefore, real household saving equals the change in the capital stock plus the change in real government bonds:

$$(14.9) \qquad \textit{real household saving} = K_t - K_{t-1} + (B_t^g - B_{t-1}^g)/P.$$

The sum of real government saving and real household saving gives us real **national saving**, that is, the total saving done by the whole nation. We can see from equations (14.3) and (14.9) that, when we add together government and household saving, the terms for the change in real government bonds, $(B_t^g - B_{t-1}^g)/P$, cancel. The reason is that an increase in real government bonds means that the government is saving less, and households are saving more. Therefore, we get

(14.10) $$\textit{real national saving} = K_t - K_{t-1}.$$

Thus, real national saving equals the change in the capital stock, which is the same as net investment. We can also note that this result continues to hold if we allow the government to own capital. In that case, we have to interpret K_t as the economy's total capital stock, that is, the sum of private and public capital. Equivalently, net investment on the right-hand side of equation (14.10) becomes the sum of private net investment and public net investment.

IV. Public Debt and Households' Budget Constraints

We found in chapter 6 that households care about the present value of real transfers net of real taxes, $V_t/P - T_t/P$. To assess the economic effects from public debt and budget deficits, we have to know how they affect this present value. To illustrate the main results, it is convenient to start with a number of simplifying assumptions. After we understand the basic result, we can generalize to more realistic situations.

A. A simple case of Ricardian equivalence

We begin with a simple case that makes a number of unrealistic assumptions, as follows:

- The real interest rate, r_t, is the same in each year. That is, $r_0 = r_1 = r_2 = \ldots = r$.
- The money stock, M_t, and the price level, P_t, are constant. In this case, the government obtains no revenue from printing money. In addition, with zero inflation, the real interest rate, r, equals the nominal rate, i.
- Real transfers, V_t/P, are zero in each year.
- The government has a given time path of purchases, G_t. That is, if the government changes its budget deficits or levels of public debt, the path of G_t stays the same.
- The government starts with no debt, so that $B_0^g = 0$.

Since real transfers, V_t/P, are zero each year, the government's budget constraint for year t simplifies from equation (14.2) to

$$G_t + r\cdot(B_{t-1}^g/P) = T_t/P + (B_t^g - B_{t-1}^g)/P. \quad (14.11)$$

Since the government starts with zero debt, we have $B_0^g/P = 0$. Therefore, in year 1, the government's real interest payments, $r\cdot(B_0^g/P)$, are zero, and the budget constraint is

$$G_1 = T_1/P + B_1^g/P \quad (14.12)$$

government purchases in year 1 = real taxes in year 1 + real debt at the end of year 1.

Suppose, to begin, that the government balances its budget each year. Then, in year 1, real spending, G_1, equals real taxes, T_1/P. In that case, equation (14.12) implies that the real public debt remains at zero at the end of year 1, that is, $B_1^g/P = 0$.

Continuing on, if the government balances its budget every year, the real public debt, B_t^g/P, is zero every year.

Suppose now that, instead of balancing its budget in year 1, the government runs a real budget deficit of one unit—that is, real spending, G_1, exceeds real taxes, T_1/P, by one unit. Since we assume no change in government purchases, G_1, the deficit must reflect a decrease by one unit in real taxes, T_1/P. Equation (14.12) implies that the real deficit of one unit requires the government to issue one unit of real public debt at the end of year 1, that is, $B_1^g/P = 1$.

Assume that the government wants to restore the public debt to zero from year 2 onward, so that $B_2^g/P = B_3^g/P = ... = 0$. We have to figure out what this policy requires for year 2's real taxes, T_2/P. To calculate T_2/P, we use the government's budget constraint for year 2. This constraint is given from equation (14.11) as

$$G_2 + r\cdot(B_1^g/P) = T_2/P + (B_2^g - B_1^g)/P \tag{14.13}$$

government purchases in year 2 + real interest payments in year 2

= real taxes in year 2 + real budget deficit in year 2.

If we substitute $B_1^g/P = 1$ and $B_2^g/P = 0$, this constraint simplifies to

$$G_2 + r = T_2/P - 1.$$

Therefore, we can rearrange terms to calculate real taxes for year 2:

$$T_2/P = G_2 + 1 + r.$$

In other words, the government must raise real taxes in year 2, T_2/P, above year 2's government purchases, G_2, by enough to pay the principal and interest, $1+r$, on the one unit of debt, B_1^g/P, issued in year 1.

Overall, year 1's real taxes, T_1/P, fall by one unit, and year 2's real taxes, T_2/P, rise by $1+r$ units. The question is, how do these changes affect the total present value of real taxes paid by households? Recall from chapter 6 that, to calculate the present value of year 2's taxes, we have to divide the real amount paid by the discount factor, $1+r$. Therefore, the effect of the two tax changes on the total present value of real taxes is given by

decrease in year 1's real taxes + present value of increase in year 2's taxes

$$= -1 + (1 + r)/(1 + r)$$

$$= -1 + 1$$

$$= 0.$$

Hence, households experience no net change in the present value of real taxes when the government runs a budget deficit in year 1 and then pays off the debt with the necessary budget surplus in year 2.

We mentioned that income effects on households depend on the present value of real transfers net of real taxes, $V_t/P - T_t/P$. In the present case, V_t/P is zero. We also found that the budget deficit in year 1, followed by the budget surplus in year 2, did not affect the total present value of real taxes for years 1 and 2. Therefore, the present value of real taxes net of real transfers did not change. We conclude that the budget deficit in year 1, followed by the surplus in year 2, has no income effects on households' choices of consumption and labor supply in each year.

We can interpret the result as follows. Households receive one unit extra of real disposable income in year 1 because of the cut in year 1's real taxes by one unit. However, households also have $1+r$ units less of real disposable income in year 2 because of the rise in year 2's real taxes by *1+r* units. If households use the extra unit of real disposable income in year 1 to buy an extra unit of bonds, they will have just enough additional funds—*1+r* units—to pay the extra real taxes in year 2. Thus, the tax cut in year 1 provides enough resources, but no more, for households to pay the higher taxes in year 2. That is why there is no income effect. Nothing is left over to raise consumption or reduce labor supply in any year.

We can look at the results as saying that households view real taxes in year 1, T_1/P, as equivalent to a real budget deficit in year 1, $(B_1^g - B_0^g)/P$. If the government replaces a unit of real taxes by a unit of real budget deficit, households know that the present value of next year's real taxes will rise by one unit. Thus, the real budget deficit is the same as a real tax in terms of the overall present value of real taxes. This finding is the simplest version of the **Ricardian equivalence theorem** on the public debt. (The theorem is named after the famous British economist David Ricardo, who first enunciated it in the early 1800s.[8])

We can also interpret the findings in terms of household saving. The real budget deficit of one unit in year 1 means that real government saving is minus one unit. Since households do not change consumption, they place the entire extra unit of real disposable income in year 1 into bonds. Therefore, real household saving in year 1 rises by one unit. Thus, the extra real household saving exactly offsets the government's real dissaving. It follows that the sum of household and government real saving—which equals real national saving—does not change. Thus, another way to express the result is that a budget deficit does not affect real national saving.

---

[8] For discussions, see Ricardo (1846), Buchanan (1958, pp. 43-46, 114-122), and Barro (1989). O'Driscoll (1977) points out Ricardo's doubts about the empirical validity of his famous theorem.

B. Another case of Ricardian equivalence

Our basic result was that a deficit-financed cut in year 1's real taxes by one unit led to an increase by one unit in the present value of future real taxes. In our simple example, all of the higher future taxes appeared in year 2. More generally, some of the increases in real taxes would occur in other years.

To get a more general result, we can drop the assumption that the government runs enough of a budget surplus in year 2 to pay off all of the bonds issued in year 1. Assume, as before, that the government issues one unit of real debt, B_1^g/P, at the end of year 1. Recall that the government's budget constraint for year 2 is

(14.13) $$G_2 + r\cdot(B_1^g/P) = T_2/P + (B_2^g - B_1^g)/P.$$

Suppose now that, in year 2, the government does not pay off the one unit of debt, B_1^g/P, issued in year 1. Assume instead that the government carries the principal of this debt, one unit, over to year 3. Therefore, we have

$$B_2^g/P = B_1^g/P = 1.$$

If we substitute $B_1^g/P = 1$ and $B_2^g/P = 1$ into equation (14.13), we get

$$G_2 + r = T_2/P.$$

Therefore, year 2's real taxes, T_2/P, cover the interest payment, r, but not the principal, 1, for the real debt, B_1^g/P, issued in year 1. In other words, the government balances its budget in year 2: the real taxes equal the real expenditure on purchases plus the real interest payments.

If the government again balances its budget in year 3, we find by the same reasoning that year 3's real taxes, T_3/P, cover the interest payment, r, on the one unit of real debt. Similarly, if the government balances its budget every year, real taxes, T_t/P, cover each year's interest payment, r. The time profile for the changes in real taxes is therefore:

- year 1: T_1/P falls by 1,
- year 2: T_2/P rises by r,
- year 3: T_3/P rises by r,

and so on. Hence, T_t/P increases by r units in each year after the first.

Think about the sequence of higher real taxes by r units in each year. What quantity of real bonds would households need at the end of year 1 to pay these taxes? If households hold one more unit of real bonds, the real interest income in year 2 would be r units, and this income could be used to pay year 2's extra taxes. Then, if the principal of the bond—one unit—were held over to year 3, the real interest income of r units could be used to pay year 3's extra taxes. Continuing this way, we find that the real interest income each year would allow households to meet their higher real taxes each year.

What then is the present value of the increase in real taxes by r units starting in year 2? This present value must be the same as the one extra unit of real bonds in year 1 needed to pay the additional real taxes in each subsequent year. But, obviously, the present value of one unit of real bonds in year 1 is one unit. Therefore, we conclude that the present value of the additional future real taxes is one.[9]

---

[9] We can verify this answer by summing the present values:

$$\frac{r}{1+r} + \frac{r}{(1+r)^2} + \frac{r}{(1+r)^3} + \ldots = \left(\frac{r}{1+r}\right)\cdot\left[1 + \left(\frac{1}{1+r}\right) + \left(\frac{1}{1+r}\right)^2 + \ldots\right].$$

The infinite sum inside the brackets has the form of the geometric series $1 + x + x^2 + \ldots$, which equals $1/(1-x)$ if x is less than one in magnitude. In our case, $x = 1/(1+r)$. Therefore, we have

Given the results about the present value of the higher future real taxes, the overall change in the present value of real taxes comes from combining two terms:

- -1: real tax cut in year 1
- +1: present value of real tax increases in future years.

Since the sum of the two terms is zero, we conclude, as in our first example, that a deficit financed cut in year 1's real taxes leads to no change in the overall present value of real taxes. Therefore, we find again that the deficit-financed tax cut in year 1 has no income effects on households' choices of consumption and labor supply in each year.[10]

C. Ricardian equivalence more generally

We now have two examples in which a deficit financed tax cut does not affect the present value of real taxes paid by households. That is, the present value of real taxes is invariant to the budget deficit. To get this answer, we made a number of unrealistic assumptions. However, we can show that the invariance result still holds if we relax most of these assumptions.

$$\frac{r}{1+r} + \frac{r}{(1+r)^2} + \frac{r}{(1+r)^3} + \ldots = \left(\frac{r}{1+r}\right)\cdot\left[\frac{1}{1-\left(\frac{1}{1+r}\right)}\right]$$

$$= \left(\frac{r}{1+r}\right)\cdot\left(\frac{1+r}{1+r-1}\right)$$

$$= \left(\frac{r}{1+r}\right)\cdot\left(\frac{1+r}{r}\right)$$

$$= 1.$$

[10] To get this answer, we have to assume that households care about future real taxes into the indefinite future. We allow in section V.F.1 for the possibility that households consider future real taxes only over a finite horizon.

We can drop the assumption that the real interest rate, r_t, is the same in each year. If r_t varies over time, we can show that the two examples still work. That is, a deficit financed tax cut still has no impact on the present value of real taxes.

We can allow for variations in the money stock, M_t, and the price level, P_t. In this case, the main new feature is that the revenue from money creation—often called the inflation tax—should be viewed as another form of tax. That is, real taxes, T_t/P_t, have to be extended to include the inflation tax. With this extension, we still find that a deficit financed tax cut has no effect on the present value of real taxes paid by households.

We can allow for real transfers, V_t/P, in each year. In this case, the result is that a deficit financed tax cut does not affect the present value of real taxes net of real transfers, $T_t/P - V_t/P$. Thus, we still find that the budget deficit has no income effects on households' choices of consumption and labor supply.

We can allow the initial real public debt, B_0^g/P, to be greater than zero. In that case, we can still consider a deficit financed cut in real taxes for year 1. This tax cut will raise the real debt from B_0^g/P to a higher value, B_1^g/P. We can then demonstrate that the larger public debt requires future real taxes to rise in present value by the same amount as the cut in year 1's real taxes. Thus, as before, a deficit financed tax cut does not change the present value of real taxes paid by households.

We can show that differences in the size of the initial real debt, B_0^g/P, do not have income effects on households' choices of consumption and labor supply. The reason is that a higher B_0^g/P requires the government to collect a correspondingly higher present value of real taxes to finance the debt. Thus, for households, the extra assets in the form of more real government bonds, B_0^g/P, is exactly offset by a higher present value of real taxes.

We can also allow for tax cuts and budget deficits in future years, not just for year 1. These future deficits will require higher real taxes in the more distant future. In each case, the present value of the higher real taxes later on exactly offsets the present value of the preceding tax cut. Therefore, we still get no effect overall on the present value of real taxes paid by households. This conclusion holds for any time pattern of budget deficits and tax cuts.[11]

In all of these cases, we find no income effects on households' choices of consumption and labor supply. The reason we keep getting the same result is that we have held fixed the government's time path of purchases, G_t. These purchases have to be paid for at some point by levying real taxes, T_t/P. By varying its budget deficits the government can change the timing of taxes. However, the government cannot escape having to levy the taxes sometime—this conclusion is an example of the famous economic adage that there is no free lunch. If the government wants to change the real present value of taxes, it has to change the present value of its purchases, G_t.

V. Economic Effects of a Budget Deficit

We consider now what happens to the economy when the government cuts year 1's real taxes, T_1/P, and runs a budget deficit. Economists often refer to this type of change as a stimulative **fiscal policy**.

[11] What if the government has a deficit financed tax cut and finances the extra public debt forever by issuing new debt? In this case, it seems that future real taxes would never increase. However, this form of financing requires an explosive path for the public debt—it amounts to a form of chain letter in which the real debt rises at a rate that is ultimately unsustainable. Our assumption is that the government cannot carry out these types of chain letters.

We know that the budget deficit has no income effects on households' choices of consumption and labor supply. The reason is that the increase in the present value of future real taxes matches the current tax cut. There may, however, be substitution effects from the tax changes. These effects, which we explored in chapter 13, depend on the forms of the taxes that are cut today and raised in the future. We begin by assuming that taxes are lump-sum, as in chapter 12. Although this case is unrealistic, it does correspond to the one usually examined in macroeconomic textbooks. We consider more realistic types of taxes in the following sections.

A. Lump-sum taxes

Suppose that the cut in year 1's real taxes, T_1/P, and the increases in future real taxes, T_t/P, all involve lump-sum taxes. The important feature of these kinds of taxes is that they have no substitution effects on consumption and labor supply. Recall that the budget deficit also has no income effects on consumption and labor supply. We conclude that, for a given real interest rate, r, and real wage rate, w/P, a deficit financed tax cut would not affect consumption, C, and labor supply, L^s. (For convenience, we now omit the time subscripts on variables.)

A deficit financed tax cut does not affect the schedules for the marginal products of labor and capital, MPL and MPK. Therefore, the deficit does not shift the demand curve for labor (Figure 8.3) or the demand curve for capital services (Figure 8.8). Since the supply curve for labor does not shift, we find that the market-clearing real wage rate, $(w/P)^*$, and quantity of labor, L^*, are unchanged (Figure 8.3). Since the supply curve for capital services also does not shift, we find that the market-clearing real rental price,

$(R/P)^*$, and quantity of capital services, $(\kappa K)^*$, are unchanged (Figure 8.8). The fixity of R/P means, as usual, that the real interest rate, r, does not change.

Since the quantities of labor, L, and capital services, κK, do not change, we know from the production function, $Y = A \cdot F(\kappa K, L)$, that real GDP, Y, must be the same. The real GDP goes, as usual, for consumption, gross investment, and government purchases:

$$Y = C + I + G.$$

We just found that Y stayed the same. We are assuming that government purchases, G, do not change. We also know that C is unchanged—because no income or substitution effects motivate households to change C. Therefore, we must have that gross investment, I, stays the same. This result tells us that today's budget deficit will not affect future capital stocks.

We can look at the results in terms of saving. Since the budget deficit has no income or substitution effects, households do not change consumption, C. However, a cut in year 1's real taxes by one unit raises households' disposable real income by one unit. Since C is the same, households must raise year 1's real saving by one unit. Therefore, the households willingly absorb the one unit of extra bonds issued by the government to cover its budget deficit. Or, as we put it before, the increase by one unit in real household saving fully offsets the reduction by one unit in real government saving. This offset means that real national saving in year 1 does not change.

We can also consider whether a budget deficit affects the price level and the inflation rate. To carry out this analysis, we now drop our assumption that the price level, P_t, equals the constant P, and the money stock, M_t, equals the constant M. That is, we return to the analysis from chapters 10 and 11, which allows for inflation and money growth.

Recall from chapter 10 that we determine the price level, P_t, for each year from the condition that the nominal quantity of money, M_t, equal the nominal quantity demanded:

(10.1) $$M_t = P_t \cdot £(Y_t, i_t).$$

The right-hand side has the nominal demand for money, M_t^d, which equals P_t multiplied by the real demand for money, $£(Y_t, i_t)$.

We have already found that a budget deficit does not affect real GDP, Y_t, and the real interest rate, r_t, in each year. Suppose, in addition, that the deficit does not affect the nominal quantity of money, M_t, in each year. That is, the monetary authority does not respond to the budget deficit by printing more money. In this case, equation (10.1) implies that the budget deficit will not affect the price level, P_t, in each year. Therefore, the deficit will not affect each year's inflation rate, π_t. We can also see that each year's real money balances, M_t/P_t, do not change.[12]

An alternative view is that budget deficits are inflationary because they encourage the government to print more money. In this case, year 1's money stock, M_1, would rise, and year 1's price level, P_1, would increase correspondingly. If the money growth rate, μ, rises, we also get increases in current and future inflation rates, π_t. However, these inflationary effects arise only if the government responds to a budget deficit by printing more money. An empirical study of ten developed countries in the post-World War II

[12] Since the real interest rate, r_t, and the inflation rate, π_t, do not change, the nominal interest rate, i_t, does not change. This result is consistent with the fixity of the real demand for money, $£(Y_t, i_t)$, on the right-hand side of equation (10.1).

period did not find this pattern.[13] That is, for these countries, which included the United States, there was no tendency for budget deficits to raise money growth rates or inflation rates.

We have found that a deficit-financed tax cut does not stimulate the economy. In particular, real GDP, Y, gross investment, I, and the real interest rate, r, do not change. If the money stock, M, does not change, the budget deficit also does not affect the price level, P. Since these results are controversial and important, we shall want to see how modifications of the model change the conclusions. We begin by assuming more realistic forms of taxes.

B. Labor-income taxes

Suppose that, instead of lump-sum taxes, the government levies taxes on labor income. As in chapter 13, let τ_w be the marginal tax rate on labor income. Consider again a reduction in year 1's real taxes, T_1/P, financed by a budget deficit. We assume now that the fall in T_1/P is accompanied by a decline in the marginal income-tax rate, $(\tau_w)_1$.

If the path of government purchases, G_t, does not change, year 1's real deficit will require real taxes, T_t/P, to rise in future years. To keep things simple, while still bringing out the main results, assume that year 2's real taxes, T_2/P, rise by enough to pay off the extra real debt issued in year 1. Thus, real taxes do not change beyond year 2. The assumption now is that the rise in T_2/P goes along with an increase in year 2's marginal income-tax rate, $(\tau_w)_2$.

[13] See Protopapadakis and Siegel (1987).

With lump-sum taxes, year 1's real budget deficit did not affect the economy. With labor-income taxes, the results are different, because the changes in the marginal income-tax rates, $(\tau_w)_1$ and $(\tau_w)_2$, affect the labor market in years 1 and 2. Figure 14.4 shows the effects for year 1. (This figure is the same as Figure 13.5, except that we now consider a decrease, rather than an increase, in τ_w.) Figure 14.4 shows that the cut in $(\tau_w)_1$ raises labor supply in year 1. This increase in labor supply leads, when the labor market clears, to a higher quantity of labor, $(L_1)'$. The higher labor input leads to a rise in year 1's real GDP, Y_1.[14]

The effects for year 2, shown in Figure 14.5, are the reverse. The increase in $(\tau_w)_2$ lowers labor supply in year 2. This decrease in labor supply leads, when the labor market clears, to a lower quantity of labor, $(L_2)'$. The reduced labor input leads to a decrease in year 2's real GDP, Y_2.

Our main finding is that a budget deficit allows the government to change the timing of labor-income tax rates and thereby alter the timing of labor input and production. Specifically, a budget deficit that finances a cut in year 1's tax rate on labor income raises year 1's labor input, L_1, and real GDP, Y_1. However, the increased public debt requires higher future taxes. If the increases in future taxes take the form of a higher income-tax rate in year 2, the levels of labor input, L_2, and real GDP, Y_2, decline. Thus, the government's fiscal policy motivates a rearrangement of the time pattern of work and production—in this case, toward the present (year 1) and away from the future (year 2).

[14] An additional effect is that the increase in L_1 tends to raise year 1's marginal product of capital services, MPK. This change affects the market for capital services in year 1. The effect is like the one shown in Figure 13.6, except that the signs are reversed. The result is that year 1's capital utilization rate, κ_1, and real rental price of capital, $(R/P)_1$, increase. The rise in capital services, $(\kappa K)_1$, contributes to the rise in real GDP, Y_1.

The specific results depend on the timing of the tax-rate decreases and increases. For example, we could modify the case just considered to allow the higher labor-income tax rates to apply for many future years, not just for year 2. In that case, the reductions in labor input, L_t, and real GDP, Y_t, would be stretched out over many years.

C. Asset-income taxes

The example studied in the previous section assumed that taxes fell on labor income. Generally, the conclusions depend on the form of the tax-rate changes. To illustrate, we can consider another form of tax that we studied in chapter 13—a tax on asset income at the rate τ_r. Assume now that year 1's budget deficit finances a cut in taxes on asset income, thereby resulting in a decrease in year 1's asset-income tax rate, $(\tau_r)_1$. Future taxes must again increase. We consider here the case in which only the tax rate, $(\tau_r)_2$, on year 2's asset income rises.

The results in chapter 13 tell us that a cut in year 1's tax rate on asset income, $(\tau_r)_1$, raises year 1's after-tax real interest rate, $[1-(\tau_r)_1]\cdot r_1$. In response, households save more and consume less. The increase in saving leads to a rise in year 1's gross investment, I_1. That is, I_1 increases, and consumption, C_1, decreases. In year 2, the tax rate on asset income, $(\tau_r)_2$, rises. Therefore, the effects are in the direction opposite to those in year 1, and households save less and consume more. The reduction in saving leads to a decrease in year 2's gross investment, I_2. Thus, I_2 falls, and consumption, C_2, rises. We see, accordingly, that the main effects are a rearrangement of the timing of investment and consumption. Investment moves toward year 1 and away from year 2, whereas consumption moves in the opposite direction.

In our previous example, we found that changes in the timing of labor-income tax rates caused changes in the timing of labor input, *L*, and real GDP, *Y*. In contrast, in the present example, changes in the timing of asset-income tax rates cause changes in the timing of consumption, *C*, and investment, *I*. The general point is that, by running budget deficits or surpluses, the government can change the timing of various tax rates. Thereby, the government can induce changes in the timing of various dimensions of economic activity—*L*, *Y*, *C*, and *I*. In the next section, we consider whether it is a good idea for the government to induce these variations in the timing of economic activity.

D. The timing of taxes and tax-rate smoothing

We have found that budget deficits and surpluses allow the government to change the timing of tax rates. However, it would not be a good idea for the government randomly to make tax rates high in some years and low in others. These fluctuations in tax rates cause unnecessary economic distortions because they give households the wrong signals in determining how to choose the time pattern of labor, production, consumption, and investment. Fortunately, the U.S. government has not behaved in this erratic manner; rather, the public debt has typically been managed to maintain a pattern of reasonably stable tax rates over time. This behavior is called **tax-rate smoothing**. This phrase signifies that the government maintains stability in tax rates even when economic disturbances occur.

One example of tax-rate smoothing concerns the response of income-tax rates to economic fluctuations. Real government expenditures typically do not decline as much in proportion as real GDP during recessions. In fact, some transfer payments, such as unemployment compensation and family welfare, tend to rise in real terms during recessions. Therefore, to maintain a balanced budget, the government would have to raise tax rates when the economy contracts. However, instead of raising tax rates, governments typically run real budget deficits during recessions and real budget surpluses during booms.[15]

As another example, during wartime, government purchases rise substantially above normal. To maintain a balanced budget, tax rates would have to be abnormally high during wars. To avoid these unusually high wartime tax rates, governments tend to run real budget deficits during wars. In this way, the necessary increases in tax rates are spread roughly evenly over time. Tax rates rise somewhat during wartime but also increase afterward to finance the public debt built up during the war. This kind of wartime deficit financing explains much of the long-term evolution of the U.S. and British public debts, as shown in Figures 14.1 and 14.2.

"Unpleasant monetarist arithmetic"

Thomas Sargent and Neil Wallace (1981) analyzed effects from changes in the timing of the inflation tax—that is, the government's revenue from printing money. Their analysis applies especially to countries such as Argentina and Brazil, which have often relied heavily on the printing press for government revenue.

[15]Because of the automatic tendency to run budget deficits during recessions, economists try to estimate what the budget deficit would have been if the economy had been operating at a level of "full capacity" or "full employment." For discussions of the full-employment deficit, see Brown (1956) and Council of Economic Advisers, *Economic Report of the President* (1962, pp. 78-82).

Suppose that the government cuts the current money growth rate in an attempt to reduce inflation. (See the discussion of money growth and inflation in chapter 11.) Assume, however, that the government does not change its current or prospective purchases and real transfers. Assume also that the government does not change its current or prospective real taxes from the income tax or other forms of taxation. In this case, the decrease in current real revenue from printing money most correspond to an increase in the real public debt—that is, to a real budget deficit. As usual, the increased real public debt implies that the government's present value of future real revenue has to rise. However, if real taxes are fixed, the future real revenue has to come from future money creation. In other words, the government is changing the timing of the inflation tax; less is collected now and more is collected later. The rise in future real revenue from money creation means that future money growth rates have to rise—that is, they have to be even higher than they were initially.

Because future money growth rates increase, today's reduction in money growth will be unsuccessful in the long run at reducing the inflation rate. The inflation rate will rise in the long run along with the money growth rate. Moreover, if people anticipate the higher future inflation rate, the cut in the current money growth rate may not even reduce the inflation rate in the short run. The reason is that the expectation of rising inflation rates tends to lower today's real demand for money and leads, thereby, to a rise in the inflation rate in the short run. Sargent and Wallace use this analysis to argue that a program to curb inflation by reducing money growth will be unsuccessful on its own. The program has to be part of a fiscal plan that cuts current or prospective real government expenditure or increases current or prospective real taxes.

E. Strategic budget deficits

In the U.S. history shown in Figure 14.1, the major exception to tax-rate smoothing concerns the real budget deficits run during the Reagan and Bush administrations after the end of the 1982-83 recession. For the remainder of the 1980s, and continuing through the first part of the 1990s, the real deficits raised the ratio of public debt to GDP, despite the absence of war or significant recession.[16] The debt-GDP ratio increased from 0.32 in 1983 to 0.40 at the end of Reagan's second term in 1988, then continued upward to reach 0.49 in 1993-94, the first two years of Clinton's first term.

One interpretation of these budget deficits is that Reagan wanted to halt the growth in the ratio of federal expenditure to GDP—see the numbers in Figures 12.1 and 12.2 of chapter 12. These data show an upward trend through the 1970s in the ratio of federal spending—particularly on transfer payments—to GDP. Reagan sought to curb the growth of both federal spending and taxes in relation to GDP. However, he was more successful, at least initially, in the second mission than the first—hence, the government ran real budget deficits. Some economists have argued that the presence of the budget

[16] The 1990-91 recession was mild and would not have exerted a major effect on the debt-GDP ratio.

deficits and the build up of public debt created political pressures on the U.S. Congress to curtail the growth of federal spending. Therefore, in the longer run, Reagan may have been successful in forcing the ratios of federal purchases and transfers to GDP to be smaller than they otherwise would have been.

This interpretation of the Reagan-Bush budget deficits after 1983 gave rise to a new theory called **strategic budget deficits**.[17] To see the basic idea, suppose that an administration—which could be the Reagan administration—is currently in power and favors a small government. Assume that this administration believes that it will be followed eventually by an administration that favors a large government (for example, the Clinton administration, which came into office in 1993). The question is, how can the Reagan administration influence future government officials of a different political persuasion to choose relatively low levels of government purchases and transfers? One answer is by running a budget deficit to leave behind a relatively high ratio of public debt to GDP. In that case, the financing of the large public debt makes it politically more difficult for the future government to select high levels of government purchases and transfers. Arguably, this is exactly what happened in the 1990s. Federal transfers did not rise relative to GDP (Figure 14.2), and federal purchases declined relative to GDP (Figure 14.1) because of large cuts in defense expenditure. Thus, the Reagan-Bush strategic budget deficits were probably successful. Similar reasoning may underlie part of the real budget deficits run during the younger Bush administration of 2001-04.

[17] This theory was developed by Persson and Svensson (1989) and Alesina and Perotti (1995).

F. The conventional view of a budget deficit

Our analysis of the economic effects of budget deficits differs dramatically from those in most macroeconomic textbooks. To see why, return to the case of lump-sum taxes, which are assumed in most macroeconomic models. For our analysis, the key point was that a deficit financed tax cut had no income effects on households' choices of consumption and labor supply. In contrast, the starting point for the conventional analysis is that a deficit-financed tax cut makes households feel wealthier. Hence, the tax cut has a positive income effect on consumption and a negative income effect on labor supply. We shall look first at how these income effects modify our conclusions about the economic effects of budget deficits. Then we briefly review the arguments that economists have made for why deficit financed tax cuts make households feel wealthier.

In our analysis in section V.A, a deficit financed cut in year 1's real taxes, (T_1/P), had no income effects. Therefore, consumption, C_1, did not change, and households saved all of the increase in year 1's real disposable income. Since real GDP, Y_1, did not change (because the inputs of labor and capital services stayed the same) and since government purchases, G_1, stayed the same, gross investment, I_1, did not change.

The different assumption in the conventional analysis is that households feel wealthier—therefore, the tax cut has a positive income effect on consumption, C_1. Suppose, to keep things simple, that labor supply does not change. In this case, year 1's inputs of labor and capital services stay the same, and real GDP, Y_1, still does not change. In this case, the increase in C_1 means that gross investment, I_1, has to decline. Thus, a major new conclusion is that a budget deficit **crowds out** investment.

Another way to look at the result is that the increase in year 1's real disposable income leads partly to more consumption and partly to more household saving. That is, household saving no longer rises by the full amount of the tax cut. Thus, national saving falls, and this decrease corresponds to the fall in gross investment, I_1.

The longer run effects depend on whether the government pays off the extra real public debt in year 2 or, instead, allows the debt to remain permanently higher. As an example of the first situation, we assumed in one case that year 2's real taxes, T_2/P, rose by enough to pay off the extra public debt. As an example of the second situation, we assumed in another case that real taxes in each year, T_t/P, rose only by enough to pay the added interest expense each year. In the latter case, the decline in investment means that the capital stock, K, will be lower in the long run than it otherwise would have been.[18] Therefore, future levels of capital services, κK, and, hence, real GDP, Y, will be lower than they would have been. Thus, additional real public debt contracts the economy in the long run.[19] These long-term negative effects on capital stock and real GDP are sometimes described as a **burden of the public debt**.[20]

Another long-run effect is that the smaller stock of capital, K, implies a higher marginal product of capital services, MPK. This higher MPK leads to a higher real rental price, R/P, which implies a higher real interest rate, r. Therefore, the conventional analysis predicts that a larger ratio of public debt to GDP leads, in the long run, to a higher r.

---

[18] In the former case, the economy returns to the original level of real public debt in year 2. Hence, there are no long-term effects on the capital stock.

[19] We can also allow for a negative income effect on labor supply. That is, households work less if the tax cut makes them feel wealthier. In this case, labor input declines, and this change reinforces the tendency for real GDP to fall.

[20] For discussions, see the papers in the volume edited by Ferguson (1964). Note especially the paper by Franco Modigliani, "Long-Run Implications of Alternative Fiscal Policies and the Burden of the National Debt."

To reach the conventional conclusions about the effects of budget deficits, we have to assume that a tax cut makes households feel wealthier. We explore next two of the more interesting justifications for this assumption. The first one concerns the finiteness of life, and the second involves imperfections of credit markets.

1. Finite lifetimes. Suppose, again, that the government runs a budget deficit and cuts year 1's real taxes, (T_1/P), by one unit. We know that the government has higher future expenses for interest and principal on the extra public debt. We also know that the real present value of the higher real taxes needed to finance these expenses is one unit, the same as the initial tax cut. Suppose, however, that some of the future taxes show up in the distant future—after the death of the typical person alive in year 1. In that case, the present value of the future taxes that accrue during the lifetimes of people living in year 1 falls short of one unit. Hence, these people experience a reduction in the present value of their real taxes.

Why does a budget deficit make people feel wealthier when they have finite lifetimes? The reason is that the decrease in the present value of real taxes for current generations coincides with an increase in the present value of real taxes for members of future generations. Individuals will be born with a liability for a portion of taxes to pay the interest and principal on the higher stock of real public debt. However, these people will not share in the benefits from the earlier tax cut. Present taxpayers would not feel wealthier if they counted fully the present value of the prospective taxes on descendants.

Budget deficits effectively enable members of current generations to die in a state of insolvency where they leave debts—that is, public debts—for their descendants.

Therefore, budget deficits make people feel wealthier if they view this governmental shifting of incomes across generations as desirable. However, most people already have opportunities for intergenerational transfers, which they have chosen to exercise to a desired extent. For example, parents make contributions to their children in the forms of educational investments, other expenses in the home, and bequests. In the other direction—and especially before the growth of social security programs—children provide support for their aged parents. To the extent that private transfers of this sort are operative, the government's budget deficit does not offer the typical person a new opportunity to extract funds from his or her descendants. Therefore, the predicted response to a rise in public debt would be a shift in private transfers by the amount necessary to restore the balance of incomes across generations that was previously deemed to be optimal. In this case, a budget deficit would not make current generations of households feel wealthier.[21] Therefore, we would get back to the results where budget deficits had no income effects on consumption and labor supply.

As a concrete example, assume that a married couple plans to leave a bequest with a present value of \$50,000 for their children. Then suppose that the government runs a budget deficit, which cuts the present value of the couple's taxes by \$1000 but raises the present value of their children's taxes by \$1000. Our prediction is that the parents use the tax cut to raise the present value of their intergenerational transfers to the children to \$51,000. The extra \$1000 provides the children with just enough resources to

[21] For a discussion of the interplay between public debt and private intergenerational transfers, see Barro (1974). A different view is that parents use bequests to control their children's behavior, rather than purely for altruistic reasons. For a discussion of this "strategic bequest theory," see Bernheim, Shleifer, and Summers (1985).

pay their higher taxes. Parents and children then end up with the same amounts of consumption that they enjoyed before the government ran its budget deficit.

2. Imperfect credit markets. Thus far, we assumed that the real interest rate, r, on private bonds equaled the rate on government bonds. Since households could issue private bonds, as well as hold them, our model assumes that households can borrow at the same real interest rate, r, as the government. In practice, however, credit markets are not this perfect. Many households who would like to borrow have to pay substantially higher real interest rates than the government. The borrowing rate is especially high if people borrow without collateral, such as a house or car.

When credit markets are imperfect, some households will calculate present values of future real taxes by using a real interest rate that is higher than the government's rate. We found before that a deficit financed cut in year 1's real taxes by one unit led to an increase in the present value of future real taxes by one unit. However, we got this result when we calculated present values by using the government's real interest rate, r. For households that face a higher real interest rate, the present value of the future real taxes will fall short of one unit.

To illustrate, suppose that the government cuts year 1's real taxes by one unit and correspondingly runs a budget deficit of one unit. Assume, as in one of our previous cases, that the government raises real taxes in year 2 by enough to pay the principal and interest on the one unit of new public debt. If the government's real interest rate is 2%, real taxes in year 2 rise by 1.02 units. To calculate the present value of these taxes, households discount the 1.02 in accordance with the real interest rate that they pay. If households use a real interest rate of 5%, the present value is

$$\textit{present value of increase in year 2's real taxes} = 1.02/1.05$$

$$\approx 0.97.$$

Thus, the overall change in the present value of real taxes is

$$\textit{change in present value of real taxes} = \textit{tax cut in year 1} + \textit{present value of increase in year 2's real taxes}$$

$$= -1 + 0.97$$

$$= -0.03.$$

Hence, the tax cut by one unit in year 1 decreases the overall present value of real taxes by 0.03 units. The effect would be larger if the government delayed its repayment of public debt beyond year 2.

Suppose now that some households (or businesses) have good access to credit and therefore use a real interest rate close to the government's rate to calculate present values of future real taxes. For these households, a deficit financed tax cut still leaves unchanged the overall present value of real taxes. So, what happens if the economy consists partly of households who face the same real interest rate as the government and partly of households who face higher real interest rates? In this case, a deficit financed tax cut leaves unchanged the present value of real taxes for the first group and reduces the present value for the second group. Thus, in the aggregate, the tax cut makes households feel wealthier.

Why does the imperfection of credit markets lead to the conclusion that households feel wealthier in the aggregate when the government runs a budget deficit?

By running a deficit, the government effectively loans money to households—the loan is one unit if real taxes fall in year 1 by one unit. Then the government effectively collects on the loan in future years when it raises real taxes. The real interest rate charged on these loans is implicitly the rate paid by the government on its bonds. Households view this loan as a good deal if the government's real interest rate is less than the rate at which households can borrow directly. That is why the overall real present value of taxes falls for households that use a high real interest rate to calculate present values.

The implicit assumption is that the government's use of the tax system is a relatively efficient way to lend to some households. That is, the government is better than financial institutions such as banks at lending funds (by cutting taxes) and then collecting on these loans in the future (by raising taxes later). If the government is really superior at this lending process, the economy will function more efficiently if the government provides more credit—that is, in our case, if the government runs a larger budget deficit. By operating more efficiently, we mean that resources will be better channeled toward higher priority uses. These uses might be for year 1's consumption or investment by households who previously lacked good access to credit.

In the end, the imperfection of credit markets can give us a reason why budget deficits affect the economy. However, the results do not resemble those from the conventional analysis, in which a larger public debt leads in the long run to lower levels of the capital stock and real GDP. With imperfect credit markets, budget deficits matter if they improve the allocation of credit, that is, if they alleviate some of the imperfections in private credit markets. Thus, in this line of argument, budget deficits matter but in a desirable way. Therefore, we cannot use this reasoning to argue that budget deficits and public debt are a burden on the economy.

Empirical evidence on the macroeconomic effects of budget deficits

An important prediction from the conventional analysis is that real budget deficits raise consumption and reduce national saving and investment. Over time, the reduction in investment leads to a lower stock of capital. This smaller capital stock implies a higher marginal product of capital services, MPK, which leads to a higher real interest rate, r.

There is no question that most government officials and news reporters, as well as many economists, believe that budget deficits reduce national saving and investment and raise real interest rates. Nevertheless, this belief is not supported by much empirical evidence. For example, Charles Plosser (1982, 1987) and Paul Evans (1987a, 1987b) carried out detailed statistical analyses of the effects of budget deficits on interest rates in the United States and other developed countries. Their main finding was that budget deficits had no significant effects on real or nominal interest rates.

Despite many empirical studies for the United States and other countries, it has proved difficult to reach definitive conclusions about the effects of budget deficits on consumption, national saving, and investment. One difficulty concerns the direction of causation. As discussed in section V.D, budget deficits often arise as responses to economic fluctuations and temporary government expenditure. Since economic fluctuations and government expenditure typically accompany changes in consumption, national saving, and investment, it is hard to distinguish the effects of budget deficits on the economy from the reverse effects.

An empirical study by Chris Carroll and Lawrence Summers (1987) avoids some of these problems by comparing saving rates in the United States and Canada. The private saving rates were similar in the two countries until the early 1970s but then diverged; for 1983-85 (the final years in their study), the Canadian rate was higher by six percentage points. After holding fixed the influences of macroeconomic variables and tax systems, Carroll and Summers isolated a roughly one-to-one, positive effect of budget deficits on private saving. This result accords with the conclusions about budget deficits that we reached with our model in section V.A. In that model, budget deficits did not affect national saving.

The Israeli experience from 1983 to 1987 comes close to providing a natural experiment to study the interplay between budget deficits and saving. In 1983, the national saving rate of 13% corresponded to a private saving rate of 17% and a public saving rate of -4%. (In this context, real public saving equals public investment less the real budget deficit.) In 1984, the dramatic rise in the budget deficit reduced the public saving rate to -11%. The interesting observation is that the private saving rate rose to 26%, so that the national saving rate changed little, actually rising from 13% to 15%. Then a stabilization program in 1985 eliminated the budget deficit, so that the public saving rate rose to 0% in 1985-86 and -2% in 1987. The private saving rate declined dramatically at the same time—to 19% in 1985 and 14% in 1986-87. Therefore, the national saving rate remained relatively stable, going from 15% in 1984 to 18% in 1985, 14% in 1986, and 12% in 1987. This dramatic episode provides an example in which the changes in private saving roughly offset the fluctuations in public saving and, therefore, led to near stability in national saving. This experience therefore accords with the conclusions from the model in section V.A.

VI. Social Security

Retirement benefits paid through social security programs are substantial in the United States and most other developed countries. Some economists, such as Martin Feldstein (1974), argue that these public pension programs reduce saving and investment. We can use our model to examine this idea.

The argument for an effect on saving applies when a social security system is not **fully funded**. In a funded setup, workers' payments accumulate in a trust fund, which provides later for retirement benefits. The alternative is a **pay-as-you-go system**, in which benefits to old persons are financed by taxes on the currently young. In this case, people who are at or near retirement age when the program begins or expands receive benefits without paying a comparable present value of taxes. Correspondingly, members of later generations (including most readers of this book) pay taxes that exceed their expected benefits in present-value terms.

The U.S. system, like that of most countries, operates mainly on a pay-as-you-go basis.[22] Although the initial plan in 1935 envisioned an important role for the social security trust fund, the system evolved after 1939 toward primarily a pay-as-you-go operation. Retirees increasingly received benefits that exceeded their prior contributions in present-value terms.

Consider the economic effects of social security in a pay-as-you-go system. We focus here on income effects and neglect the types of substitution effects from taxes and transfers that we discussed in chapter 13.

The usual argument goes as follows. When a social security system starts or expands, old persons experience an increase in the present value of their social security benefits net of taxes. That is, the present value of real transfers net of real taxes rises for the currently old. This change implies a positive income effect on the consumption of old persons.

Young persons face higher taxes, offset by the prospect of higher retirement benefits. Thus, the present value of real transfers net of real taxes may fall for the currently young. However, this decline is not as large in magnitude as the increase for

[22] Privatized arrangements, such as the main pension system in Chile, are fully funded but do not involve a government trust fund. For a discussion, see Pinera (1996). The World Bank (1994) provides an overview of social security systems throughout the world.

the currently old. Why? Because the currently young will be able to finance their future retirement benefits by levying taxes on yet unborn future generations. Thus, although there may be a negative income effect on the consumption of young persons, we predict that this effect will be smaller in size than that for the currently old. Hence, we predict an increase in current aggregate consumption. Or, to put it another way, aggregate household saving declines. This change leads in the short run to a decrease in investment and, in the long run, to a reduced stock of capital.

This analysis of the economic effects of social security parallels the conventional analysis of a budget deficit considered in section V.F. In both cases, the increase in aggregate consumption arises only if people neglect the adverse effects on descendants. Specifically, an increase in the scale of a pay-as-you-go social security program means that the typical person's descendants will be born with a tax liability that exceeds his or her prospective retirement benefits in present-value terms. If the people who are currently alive take full account of these effects on descendants, the income effects from a social-security program would be nil.

As in the case of a deficit-financed tax cut, more social security enables older persons to extract funds from their descendants. However, also as before, people value this change only if they give no transfers to their children and receive nothing from their children. Otherwise, people would respond to more social security by shifting private intergenerational transfers, rather than by consuming more. In the United States, for example, the growth of social security has strongly diminished the tendency of children to support their aged parents.

On an empirical level, there has been a great debate since the 1970s about the connection of social security to saving and investment. Martin Feldstein (1974) reported a dramatic negative effect of social security on capital accumulation in the United States. However, subsequent investigators argued that this conclusion was unwarranted.[23] Neither the long-term evidence for the United States nor that from a cross section of countries in recent years provides convincing evidence that social security depresses saving and investment.

VII. Open-Market Operations

The inclusion of public debt in the model allows us to analyze **open-market operations**. An open-market purchase occurs when the central bank, such as the Federal Reserve, buys bonds—typically government bonds—with newly created money. An open-market sale occurs when the central bank sells bonds for money. These open-market operations are the main way that the Federal Reserve and most other central banks control the quantity of money. We shall want to see whether this realistic way of changing the quantity of money leads to results that differ from those for the unrealistic "helicopter drops" of money that we studied in chapters 10 and 11.

Consider an open-market purchase, whereby the quantity of money, M, increases by \$1, and the stock of government bonds, B^g, decreases by \$1. Assume that no subsequent changes in the quantity of money occur; that is, we are considering a one-time increase in M.

Table 14.2 shows that an open-market purchase of government bonds amounts to a combination of two policies that we have already studied. Suppose, first, that the

[23] For a summary of the debate, see Esposito (1978) and Leimer and Lesnoy (1982).

government prints an extra dollar of money, M, and uses the money to cut lump-sum taxes by \$1 or raise lump-sum transfers by \$1. For example, the government might have a helicopter drop of \$1 of money—the unrealistic story considered in chapters 10 and 11. This change is labeled as policy 1 in the table. Suppose, next, that the government raises lump-sum taxes by \$1 or cuts lump-sum transfers by \$1 and uses the proceeds to pay off \$1 of government bonds, B^g. That is, the government raises taxes net of transfers and, thereby, runs a budget surplus. These changes are called policy 2 in the table. If we combine the two policies, we find that the quantity of money, M, rises by \$1, taxes and transfers are unchanged, and the quantity of government bonds, B^g, falls by \$1. Thus, we end up with an open-market purchase of government bonds, shown as policy 3 in the table.

We know from chapter 10 that policy 1—a one-time increase in the quantity of money, M, used to cut lump-sum taxes or raise lump-sum transfers—raises the price level, P, in the same proportion as the increase in M. We also know that there are no effects on real variables, including real GDP, Y, the real interest rate, r, and the real quantity of money, M/P.[24] We know from our analysis in this chapter that policy 2—the budget surplus created by an increase in taxes net of transfers—does not affect the same group of real variables, including Y, r, and M/P. For a given nominal quantity of money, M, the budget surplus also does not affect the price level, P. Therefore, the overall effect from policy 3—the open-market purchase of government bonds—is that P rises in the same proportion as M, and the group of real variables does not change. That is, an open-market purchase has the same effects as the unrealistic helicopter drop of

[24] With government bonds in the model, one new effect is that the real quantity of these bonds, B^g/P, falls because of the increase in the price level, P. However, in the model, changes in B^g/P do not affect any other real variables, such as real GDP, Y, or the real interest rate, r.

money that we considered in chapters 10 and 11. We conclude that this unrealistic story gave us a reasonable and simple way to assess the linkages between money and the price level.

VIII. Summing Up

We allowed the government to borrow, that is, to run a budget deficit by issuing bonds. The government's budget constraint then includes two new items: interest payments on the existing public debt and the proceeds from new debt issue.

In our basic model, changes in budget deficits have no income effects. The reason is that, if the paths of government purchases and transfers do not change, a deficit-financed tax cut creates future taxes with the same present value as the current tax cut. If we assume lump-sum taxes, the absence of an income effect implies that variations in budget deficits have no effect on real GDP, consumption, investment, and the real interest rate. This result is known as Ricardian equivalence—taxes and budget deficits have equivalent effects on the economy. For labor-income or asset-income taxes, variations in budget deficits affect the timing of distorting tax rates and, thereby, influence the time paths of labor, real GDP, consumption, and investment. In most circumstances, governments avoid these timing effects by following a policy of tax-rate smoothing. This policy attempts to keep tax rates stable despite economic fluctuations and variations in government expenditure. To smooth tax rates, the government runs budget deficits during recessions and wars. One deviation from tax-rate smoothing, known as strategic budget deficits, represents the attempt by an administration to use budget deficits to influence future choices of government expenditure. The idea is that, by leaving behind a larger stock of public debt, a government can make it more difficult for a future administration to choose high levels of purchases and transfers.

The conventional view of budget deficits assumes that a deficit-financed tax cut has a positive income effect on consumption. In this case, a budget deficit tends to raise consumption and lower investment. In the long run, a larger public debt leads to a lower capital stock and a higher real interest rate. This view of budget deficits is often rationalized by the finiteness of life and the imperfections of private credit markets.

The inclusion of government bonds allows us to analyze pay-as-you-go social security systems and open-market operations by the monetary authority. An expansion of social security is analogous to a deficit-financed tax cut. Open-market operations have effects similar to those from the unrealistic helicopter drops of money considered in previous chapters.

Questions and Problems

Mainly for review

14.1. Under what circumstances is an open-market operation neutral?

14.2. Suppose that the government announces a reduction in future tax rates on labor income. What intertemporal-substitution effect will this announcement have on current labor supply?

Problems for discussion

14.x. The income effect from a budget deficit

Suppose that, in year 1, the government cuts current, lump-sum taxes and runs a budget deficit. Assume that the real public debt remains constant in future years. Also, no changes occur in government purchases of goods and services, G, or in real transfers, V/P. Analyze the income effect from the government's tax cut. How does this effect depend on the following:

a. finite lifetimes?

b. the existence of childless persons?

c. uncertainty about who will pay future taxes?

d. the possibility that the government will print more money in the future rather than levying future taxes?

e. the imperfection of private credit markets?

14.x. The Reagan tax-cut plan in 1981

President Reagan's initial proposal in 1981 for cutting U.S. individual income-tax rates involved a 23% overall reduction in rates. The full cut was to be phased in over a three-year period ending in 1983. The plan also involved gradual reductions in real government expenditure when expressed as a fraction of real GDP.

Consider an alternative plan that would have yielded the same present value of real tax revenue but that implemented the entire cut in tax rates in 1981. Assume that real government expenditure is the same as under Reagan's plan. Compare this plan with Reagan's with respect to the effects on labor and real GDP over the period 1981-83.

Similarly, the Bush tax cut of 2001 scheduled decreases in marginal income-tax rates from 2001 to 2006. The 2003 law advanced the scheduled tax-rate cuts to 2003. How does your analysis of the Reagan tax plan apply to the 2001 and 2003 tax cuts?

14.x. Social security and the capital stock

Suppose that the government introduces a new social-security program, which will make payments to covered persons when they retire.

a. What long-run effects do you predict on the capital stock, K?

b. How does your answer to part a. depend on whether the social-security program is fully funded or pay-as-you-go? (In a funded scheme, workers pay into a trust fund, which is then used to pay benefits. A pay-as-you-go system taxes current workers to pay benefits for current retirees.)

Table 14.1

Public Debt in the United States and the United Kingdom

Year	U.S. Debt	U.S. Debt-GDP ratio	U.K. Debt	U.K. Debt-GDP ratio
	($ billion)		(₤ billion)	
1700	--	--	0.014	0.20
1710	--	--	0.030	0.31
1720	--	--	0.039	0.49
1730	--	--	0.037	0.49
1740	--	--	0.033	0.40
1750	--	--	0.059	0.68
1760	--	--	0.083	0.79
1770	--	--	0.106	1.01
1780	--	--	0.135	1.01
1790	0.075	0.33	0.179	1.05
1800	0.083	0.20	0.304	0.79
1810	0.048	0.082	0.436	0.96
1820	0.090	0.12	0.568	1.37
1830	0.039	0.038	0.544	1.12
1840	0.005	0.003	0.562	1.01
1850	0.063	0.029	0.557	0.94
1860	0.065	0.017	0.589	0.69
1870	2.04	0.28	0.593	0.51
1880	1.71	0.15	0.591	0.43
1890	0.711	0.054	0.578	0.37
1900	1.02	0.054	0.628	0.31
1910	0.913	0.028	0.665	0.28
1920	23.3	0.26	7.62	1.22
1930	14.8	0.16	7.58	1.55
1940	42.8	0.42	10.5	1.37
1950	219	0.74	26.1	1.77
1960	237	0.44	28.4	1.09
1970	283	0.27	33.4	0.64
1980	712	0.25	113	0.49
1990	2412	0.42	190	0.35
2000	3438	0.35	369	0.39
2001	3339	0.33	370	0.37
2002	3553	0.34	391	0.37
2003	3924	0.36	--	--
2004	4320	0.37	--	--

Notes to Table 14.1

For the United States, the public debt in billions of dollars is the amount of privately held, interest-bearing debt of the U.S. federal government at nominal par value. The figures are net of holdings by the Federal Reserve and U.S. government agencies and trust funds. (They include holdings by government sponsored agencies and by state and local governments.) For the sources, see Barro (1978a, Table 1). Values since 1939 are from U.S. Treasury Department. The debt is expressed as a ratio to nominal GDP or GNP. The data on real GDP or GNP are described in Figure 1.1 of chapter 1. Before 1869, nominal GNP is estimated by multiplying real GNP by a price index based on wholesale prices from Warren and Pearson (1933, Table 1).

For the United Kingdom, the public debt since 1917 is the central government's gross sterling debt in billions of pounds at nominal par value. Before 1917, the figures represent the cumulation of the central government's budget deficit, starting from a benchmark stock of public debt in 1700. For discussions of the data, see Barro (1987). The underlying data are from Central Statistical Office, *Annual Abstract of Statistics*, various issues; Mitchell and Deane (1962); and Mitchell and Jones (1971). Data on nominal GDP or GNP are from the preceding and also from Feinstein (1972) and Deane and Cole (1969). Before 1830, nominal GNP is estimated from rough estimates of real GNP multiplied by a price index based on wholesale prices.

Table 14.2

Open-Market Purchases of Government Bonds

Government policy	Change in money, M	Change in government bonds, B^g	Change in taxes, T
1. Print more money and reduce taxes.	+ \$1	0	- \$1
2. Raise taxes and reduce public debt.	0	- \$1	+ \$1
3. Open-market purchase of government bonds.	+ \$1	- \$1	0

Note: Policy 3—an open-market purchase of government bonds—amounts to a combination of policies 1 and 2, which we have already studied.

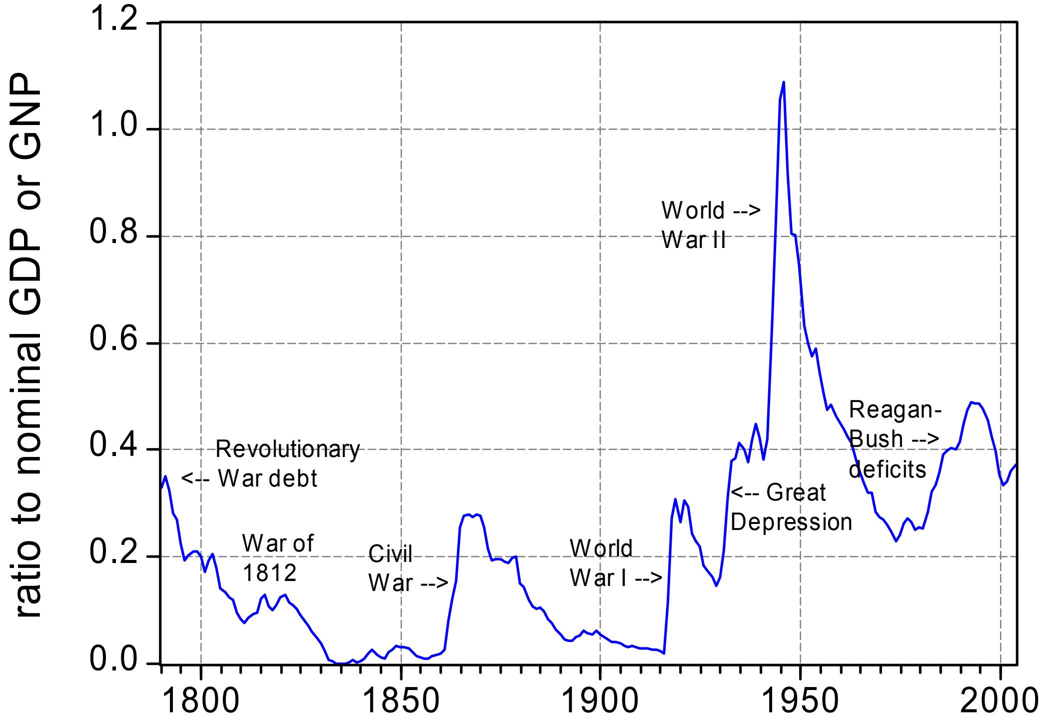


Figure 14.1

Ratio of U.S. Public Debt to GDP, 1790-2004

The graph shows the ratio of U.S. nominal public debt to nominal GDP (GNP before 1929), using the numbers described in Table 14.1.

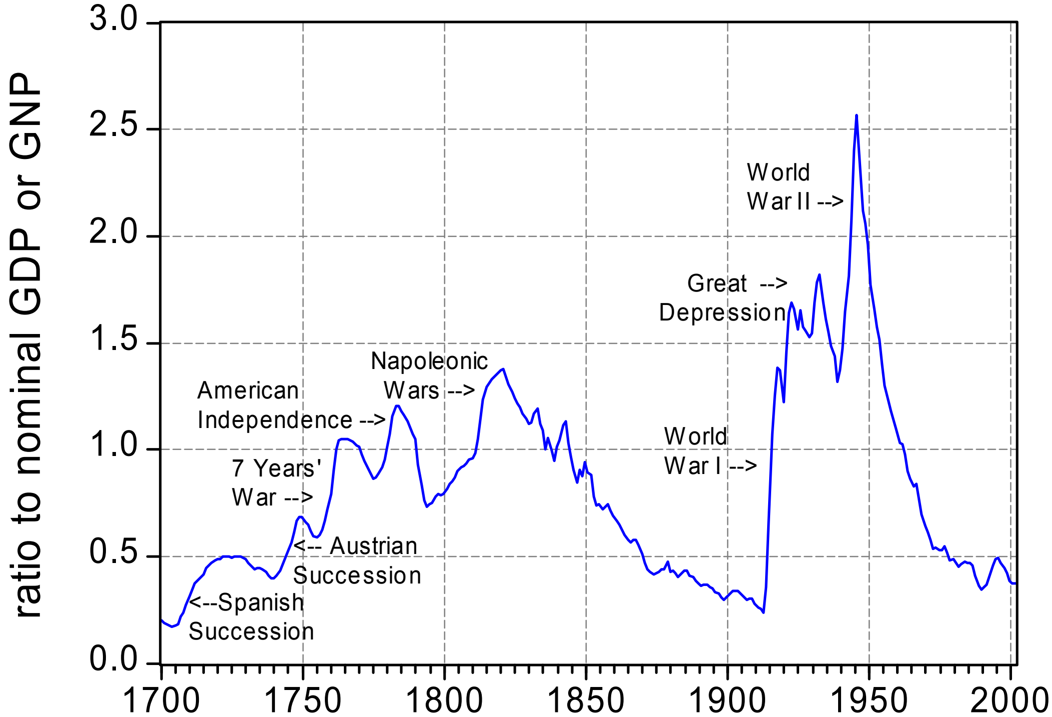


Figure 14.2

Ratio of U.K. Public Debt to GDP, 1700-2002

The graph shows the ratio of U.K. nominal public debt to nominal GDP or GNP, using the numbers in Table 14.1.

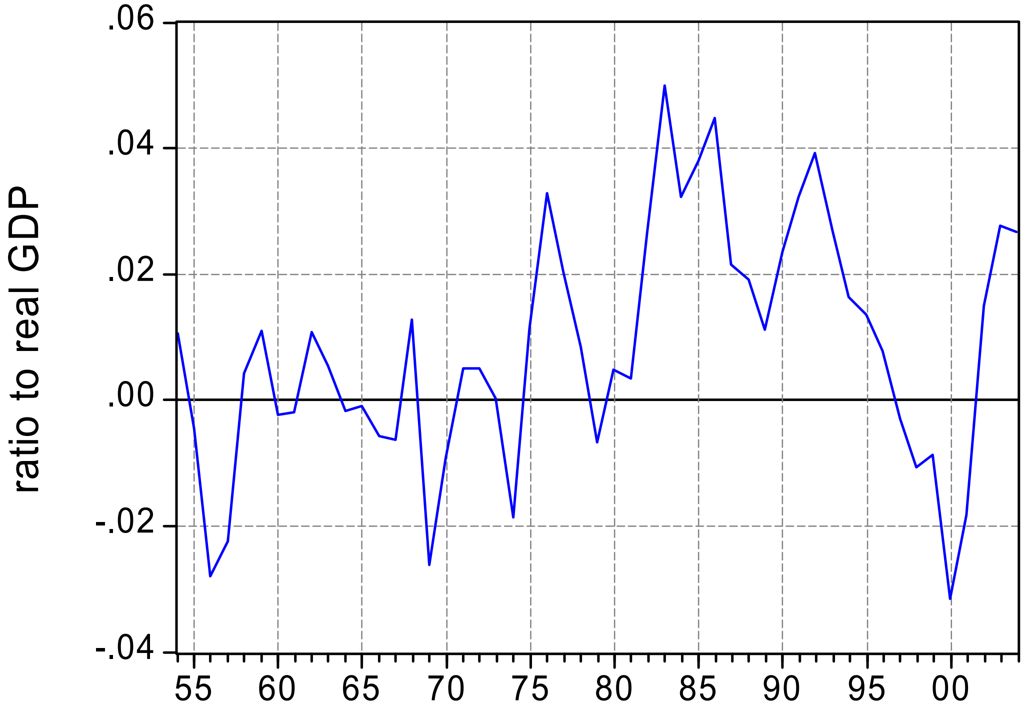


Figure 14.3

Ratio of U.S. Real Budget Deficit to Real GDP

We calculate the real public debt for each year as the ratio of the nominal debt to the GDP deflator. The real budget deficit is the change each year in the real public debt. The graph shows the ratio of the real budget deficit (the change in the real public debt) to real GDP.

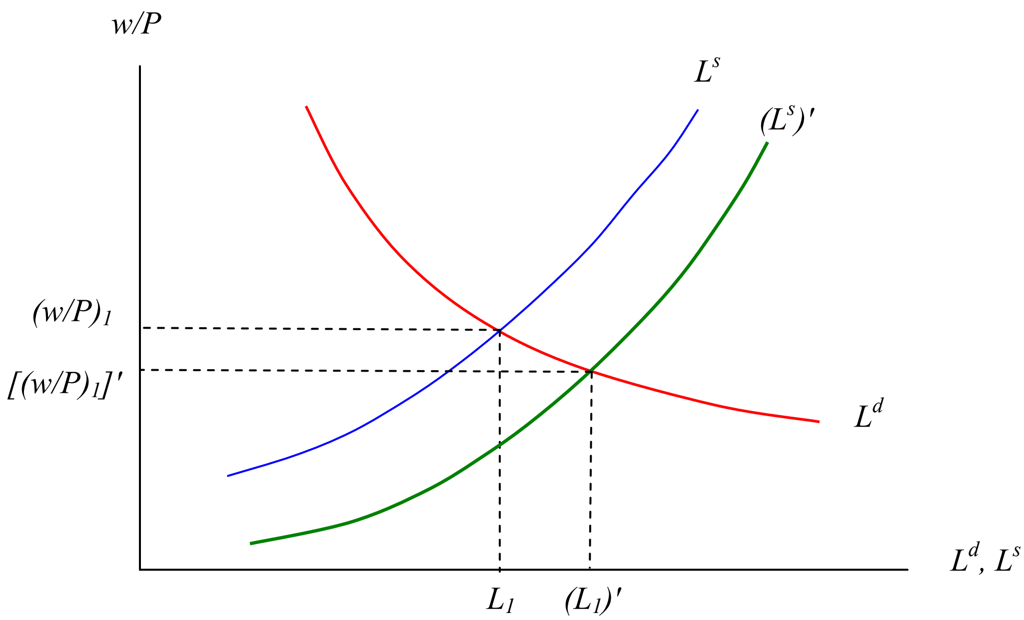


Figure 14.4

Effect of a Decrease in Year 1's Labor-Income Tax Rate on the Labor Market

The downward-sloping labor-demand curve, L^d, shown in red, comes from Figure 12.6. The upward-sloping labor-supply curve, L^s, shown in blue, also comes from Figure 12.6. A decrease in year 1's marginal tax rate on labor income, $(\tau_w)_1$, shifts the labor-supply curve rightward to the green one, $(L^s)'$. Consequently, year 1's market-clearing before-tax real wage rate falls from $(w/P)_1$ to $[(w/P)_1]'$. The market-clearing quantity of labor rises from L_1 to $(L_1)'$.

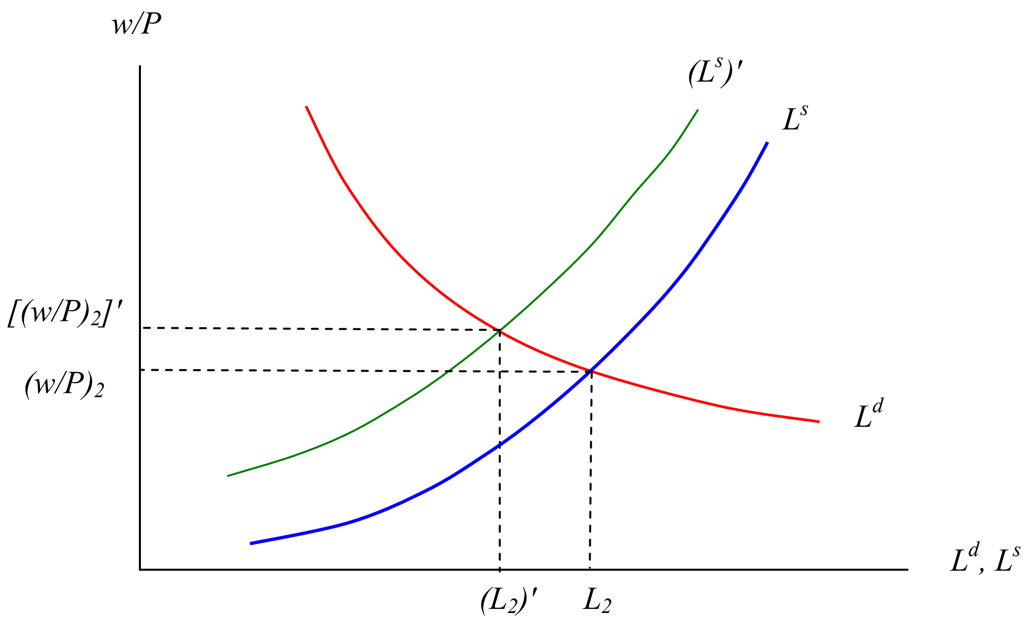


Figure 14.5

Effect of an Increase in Year 2's Labor-Income Tax Rate on the Labor Market

The downward-sloping labor-demand curve, L^d, shown in red, comes from Figure 12.6. The upward-sloping labor-supply curve, L^s, shown in blue, also comes from Figure 12.6. An increase in year 2's marginal tax rate on labor income, $(\tau_w)_2$, shifts the labor-supply curve leftward to the green one, $(L^s)'$. Consequently, the market-clearing before-tax real wage rate rises from $(w/P)_2$ to $[(w/P)_2]'$. The market-clearing quantity of labor falls from L_2 to $(L_2)'$.

January 12, 2005

Chapter 16

Money and Business Cycles II:

Sticky Prices and Nominal Wage Rates

In the previous chapter, we developed a model where monetary shocks were non-neutral because of misperceptions about prices. In that model, the price level and nominal wage rate adjusted rapidly to balance the quantities of goods or labor supplied and demanded in each market. That is, we continued to assume clearing of all the markets in the economy. By affecting labor supply, monetary shocks influenced the market-clearing quantities of labor and output and were, therefore, non-neutral.

In this chapter, we work out another type of model that economists have developed to explain non-neutrality of money. In these models, the price level and nominal wage rate are sticky—they do not adjust instantaneously to clear all the markets.

John Maynard Keynes (1936), in his *General Theory*, stressed the importance of stickiness in nominal prices and wage rates. His emphasis was on the stickiness of nominal wage rates. He thought it was unrealistic to assume that nominal wage rates adjusted rapidly to ensure continual balance between the quantities of labor supplied and demanded. Modern economists—in a field called **New Keynesian Economics**—focus instead on stickiness of nominal goods prices. As we shall see, the reason for the new focus is that it fits better in some respects with observed features of economic fluctuations. We begin our analysis by developing a model with sticky nominal goods prices.

I. The New Keynesian Model—Monopolistic Competition in the Goods Market

The usual explanation for sticky nominal goods prices relies on two main ingredients. First, the typical producer actively sets the price of the good that he or she sells in the market. This price-setting behavior differs from our previous analysis, where each perfectly competitive producer (a business or household) took the price as given by the market.   Second, when choosing the price to set, each producer takes into account a cost of changing prices. This cost is sometimes called a **menu cost**, analogous to the expense that a restaurant incurs when it alters the prices listed on its menu.

Our previous setting of perfect competition applies most naturally to large-scale organized markets on which standardized goods are traded. Examples are stock exchanges (on which financial claims are traded) and commodity exchanges (on which claims to goods such as oil or corn are traded). In these organized markets, each trader takes as given the market prices of goods. That is, each participant is small enough to neglect the impact of his or her actions on the market price.

The situation is different for markets with small numbers of sellers and buyers. For example, the markets for automobiles or computers in the United States have relatively few producers. Moreover, the goods traded on these markets are not fully standardized. Each kind of automobile or computer has different features, some of which are distinguished by brand names. These kinds of markets have substantial competition, but not the perfect competition associated with price-taking behavior by all participants.

Instead, each producer has some latitude in deciding what price to set. Economists call this environment **imperfect competition**.

In a perfectly-competitive market, a producer who charges more than the market price would find that the quantity demanded would fall to zero. Conversely, a producer who asks less than the market price would find that the quantity demanded would soar toward infinity. In markets with imperfect competition, a decrease in a seller's price generates a finite increase in the quantity of goods demanded from that seller. Similarly, a rise in price leads to a finite reduction in the quantity demanded. Thus, each producer can make a meaningful choice of what price to charge. This perspective applies to many large companies, such as automobile producers, but also holds for small businesses, such as neighborhood grocery stores. One reason that a retail store, such as a grocery, has some latitude in setting its price is that its location is convenient for buyers who live nearby or are familiar with the store. For this reason, a small increase in price above that offered by other stores would not, at least immediately, drive the quantity demanded to zero.

To illustrate the main points, we shall find it convenient to work with a formal model. Let $P(j)$ be the price charged for a good by firm j.[1] The quantity of firm j's goods demanded, $Y^d(j)$, depends on how high $P(j)$ is compared to prices charged by other producers. For example, if firm k is a competitor—perhaps because it is the grocery store in the next block—then a cut in k's price, $P(k)$, reduces $Y^d(j)$.

Generally, the quantity demanded of firm j's goods, $Y^d(j)$, will be more sensitive to prices charged on similar goods in nearby locations than to prices offered on very

[1] As in our previous models, we can think of each business as owned and run by one of the households in the economy.

different products or in faraway locations. However, the model would become unmanageable if we tried to keep track of all of these prices. We can get the main results by assuming that the customers of firm *j* compare the price, $P(j)$, with the average of the prices charged by other firms. If we let this average price be P, then $Y^d(j)$ depends on the price ratio, $P(j)/P$. An increase in $P(j)/P$ lowers $Y^d(j)$, and a decrease in $P(j)/P$ raises $Y^d(j)$.

The quantity of goods demanded at firm *j*, $Y^d(j)$, depends also on the incomes of persons who are present or potential customers of the firm. For example, if the real income in the whole economy increases, the demand, $Y^d(j)$, will rise for each firm.

The production function for firm *j* is

(16.1) $$Y(j) = F[K(j), L(j)],$$

where $K(j)$ and $L(j)$ are the quantities of capital and labor used by firm *j*. To keep things simple, we ignore changes in capital utilization, $\kappa(j)$, and we assume that $K(j)$ is fixed in the short run.[2]

Suppose that the nominal wage rate, w, is the same for labor used by all the firms in the economy. In other words, we are thinking of labor as a standardized service that is traded in the overall economy under conditions of perfect competition. More specifically, when we consider sticky nominal prices, we will ignore the possibility of sticky nominal wage rates—w adjusts to balance the total quantities of labor supplied and demanded in the economy. This assumption about the labor market can be questioned and was surely not the setup envisioned by Keynes (1936) in his *General Theory*. We allow later for sticky nominal wage rates.

[2] In addition, we are neglecting inputs of intermediate goods, which are materials and products produced by other firms. Extensions to allow for variable capital utilization and intermediate inputs would not change the main conclusions.

We begin with a setup where the nominal price charged by each firm, $P(j)$, is fully flexible. That is, we ignore any menu costs for changing prices. Given the nominal wage rate, w, and the average nominal price, P, charged by competitors, each firm sets $P(j)$ at the level that maximizes its profit.

One determinant of the profit-maximizing price, $P(j)$, is the nominal cost of producing an additional unit of goods, that is, the **marginal cost of production**. To relate this concept to our discussion of labor demand in chapter 6, recall that the marginal product of labor, MPL, is the ratio of additional output, ΔY, to additional labor input, ΔL. Therefore, for firm j:

$$MPL(j) = \Delta Y(j)/\Delta L(j). \tag{16.2}$$

If we rearrange this condition, we get that the added labor, $\Delta L(j)$, needed to raise output by $\Delta Y(j)$ units is

$$\Delta L(j) = \Delta Y(j)/MPL(j). \tag{16.3}$$

The higher $MPL(j)$ the lower the quantity of labor, $\Delta L(j)$, needed to raise output by the amount $\Delta Y(j)$. If we set $\Delta Y(j) = 1$, the labor needed to raise output by one unit is

$$\Delta L(j) = 1/MPL(j).$$

The nominal cost of each unit of labor is the nominal wage rate, w. Therefore, the added nominal cost of raising output by one unit is $w\cdot[1/MPL(j)]$. In other words, the nominal marginal cost of production for firm j is

$$\begin{aligned} &\textit{firm } j\textit{'s nominal marginal cost} = w/MPL(j) \\ &= \textit{ratio of nominal wage rate to marginal product of labor}. \end{aligned} \tag{16.4}$$

Thus, for given MPL(j), a higher w means a higher nominal marginal cost.

Under perfect competition, profit maximization dictates that each firm's nominal marginal cost, given from equation (16.4) by $w/MPL(j)$, equal its price, $P(j)$.[3] Under imperfect competition, each firm instead sets $P(j)$ above its nominal marginal cost. The ratio of $P(j)$ to the nominal marginal cost is called the **markup ratio**:

(16.5) *firm j's markup ratio = P(j)/(firm j's nominal marginal cost)*.

The markup ratio that a firm chooses depends on how sensitive a firm's product demand, $Y^d(j)$, is to $P(j)$. More market power—that is, less competition—tends to imply less sensitivity of demand to price and, therefore, motivates a higher markup ratio. If the sensitivity of demand becomes extremely high, the markup ratio approaches one. That is, we get close to the perfectly competitive environment in which $P(j)$ equals nominal marginal cost. In our analysis, we assume that each firm's markup ratio is a given constant.

We can rearrange the formula for the markup ratio to get an expression for each firm's price:

(16.6) *P(j) = (firm j's markup ratio)▪(firm j's nominal marginal cost)*.

Therefore, for a given markup ratio, an increase in firm *j*'s nominal marginal cost raises its price, $P(j)$, in the same proportion. For example, if nominal marginal cost doubles, $P(j)$ doubles. If we substitute the formula for nominal marginal cost from equation (16.4), we get

(16.7) $P(j) = (\text{markup ratio}) \cdot [w/MPL(j)]$.

Therefore, for a given markup ratio, an increase in the nominal wage rate, w, leads to a rise in the same proportion in the nominal price, $P(j)$. That is, a doubling of w causes all

[3] If we rearrange the terms, the condition is $MPL(j) = w/P(j)$. Aside from the index *j*, this equation is the same as the condition for profit maximization worked out in chapter 6 under perfect competition.

firms in the economy to double their nominal prices. Therefore, the average of these prices, *P*, also doubles.

In our model, we do not have to worry about the full array of individual prices, *P(j),* that prevails in equilibrium. The important point is that, with imperfect competition, the profit-maximizing decisions of firms determine a distribution of the *P(j)*. For example, if we think about grocery stores, some of the stores will have relatively high *P(j),* whereas others will have relatively low *P(j)*.

Each firm demands labor, and the total of these demands determines the economy-wide labor demand, L^d. As in our previous models, the equilibrium of the economy-wide labor market equates the aggregate quantity of labor demanded, L^d, to the quantity supplied, L^s. This condition determines the economy-wide real wage rate, *w/P*, as well as the total quantity of labor, *L*. Finally, if we know *w/P* and *P*, we can calculate the economy-wide nominal wage rate, *w* (by multiplying *w/P* by *P*).

A. Responses to a monetary shock

Consider what happens when a monetary shock occurs. To be concrete, imagine that the nominal quantity of money, *M*, doubles. We found in chapter 10 that this change in money was neutral. Specifically, the nominal price level, *P*, and the nominal wage rate, *w*, each doubled. This change left intact the real variables in the economy, including the quantity of real money balances, *M/P*, and the real wage rate, *w/P*.

With fully flexible prices and wages, money would still be neutral in the model that includes an array of imperfectly competitive firms. In this setting, each nominal price, *P(j)*, doubles when *M* doubles. Therefore, the average price, *P*, doubles, as in the model of chapter 10. The economy-wide nominal wage rate, w, also doubles, as before. These changes leave unchanged the real variables in the economy. The real variables now include not only the economy-wide real wage rate, w/P, but also the ratio of each firm's price to the average price, $P(j)/P$.

New results arise when we allow for stickiness in the nominal price, $P(j)$, set by each firm. As mentioned before, these prices would not change all the time because of menu costs for changing prices. To illustrate the effects of price stickiness, we can consider the extreme case in which all of the $P(j)$ are rigid in the short run. The average price, P, would then also be fixed. If P is constant and the nominal quantity of money, M, doubles, each household would have twice as much real money, M/P, as before. However, nothing has happened to motivate households to hold more money in real terms. Each household would therefore try to spend its excess money, partly by buying the goods produced by the various firms.[4] Each firm would then experience an increase in the quantity demanded of its goods, $Y^d(j)$.

How does a business react when it sees an increase in demand, $Y^d(j)$, while its price, $P(j)$, is fixed (by assumption)? We noted before that, under imperfect competition, the markup ratio is greater than one. That is, the price of goods sold, $P(j)$, is greater than nominal marginal cost. This condition means that, at a given $P(j)$, an expansion of production and sales, $Y(j)$, would raise firm j's profit. For example, if $Y(j)$ rose by one unit, the added nominal revenue would be $P(j)$, whereas the added nominal cost would be the nominal marginal cost, which is less than $P(j)$. The important result is that, if $P(j)$ is fixed, a profit-maximizing firm would—over some range—meet an increase in demand by raising production, $Y(j)$.

[4] Households might also buy interest-bearing assets, that is, bonds.

To raise its production, $Y(j)$, a firm has to increase its quantity of labor input, $L(j)$. Therefore, the quantity of labor demanded, $L^d(j)$, rises in accordance with equation (16.3):[5]

$$\Delta L^d(j) = \Delta Y(j)/MPL(j). \tag{16.3}$$

The important point is that, with a fixed price, $P(j)$, an increase in the nominal quantity of money, M, leads to an expansion of labor demand by each firm.

Consider now how an increase in money, M, affects the economy-wide labor market. Since each firm increases its labor demand, $L^d(j)$, the aggregate quantity of labor demanded, L^d, is higher at any given economy-wide real wage rate, w/P. We show this effect in Figure 16.1. The increase in the quantity of money from its initial value, M_1, to the higher value, M_2, shifts labor demand rightward from the blue curve to the red curve.

We assume, as in Figure 8.3 of chapter 8, that an increase in the real wage rate, w/P, raises the quantity of labor supplied, L^s. Thus, L^s is given by the upward-sloping green curve in Figure 16.1.

We see from Figure 16.1 that an increase in nominal money from M_1 to M_2 raises the market-clearing labor input from L_1 to L_2. With the increase in labor input, each firm produces more goods in accordance with the production function of equation (16.1). Thus, the economy's total output—the real GDP, Y—increases.[6] We therefore have that a monetary expansion is non-neutral. An increase in money expands the real economy. Moreover, labor input, L, moves in a procyclical manner—it rises along with Y.

---

[5] The expansion of $L(j)$ reduces the marginal product of labor, $MPL(j)$, and therefore raises the nominal marginal cost of production, given by equation (16.4). Hence, at a fixed price $P(j)$, a profit-maximizing firm would be willing to meet an increase in demand only up to the point at which the nominal marginal cost rose to equal $P(j)$.

[6] We have not considered that the rise in Y increases the quantity of money demanded, M^d. This change will dampen the effects from the expansion of M.

Thus far, the predictions are similar to those from the price-misperceptions model, considered in chapter 15. That model also gave the result that a monetary expansion—if it generated an unperceived increase in the price level, *P*—raised real GDP, *Y*, and labor input, *L*. However, a difference between the two models concerns the real wage rate, *w/P*. In the price-misperceptions model, an expansion of *L* had to be accompanied by a fall in *w/P* in order to induce employers to use more labor input. Thus, that model predicted—counterfactually—that *w/P* would be countercyclical. We now demonstrate that the New Keynesian model does not have this problem.

Figure 16.1 shows that a monetary expansion increases the market-clearing real wage rate from $(w/P)_1$ to $(w/P)_2$. Therefore, the model generates a procyclical pattern for *w/P*. Thus, the New Keynesian model correctly predicts that *w/P* will be procyclical. The reason that the model gets this result is that employers are willing to employ more labor even though *w/P* is higher. The margin provided by the monopoly markup means that—at fixed prices of goods—firms can profitably use more labor to produce and sell more goods even though the real cost of production has gone up. The monetary expansion cuts into the markup ratios of firms. However, as long as the markup ratio remains above one, firms are willing to expand labor input and production.

As in the real business cycle model studied in chapter 8, the prediction for procyclical labor input, *L*, in the New Keynesian model depends on the upward slope of the labor-supply curve, L^s, in Figure 16.1. That is, the analysis relies on the assumption that an increase in the real wage rate, *w/P*, motivates households to work more.

One respect in which the New Keynesian model works less well than the real business cycle model concerns the average product of labor, or labor productivity, *Y/L*.

In the real business cycle model, we found in chapter 8 that Y/L was procyclical because of the direct effect of a change in the technology level, A, on the production function. Because of the increase in A, Y/L rose in a boom even though an increase in L leads to diminishing average product of labor.[7] The prediction for procyclical labor productivity accords with the U.S. evidence, summarized in Figures 8.10 and 8.11 of chapter 8.

In contrast, the New Keynesian model assumes that the technology level, A, is fixed. In addition, we did not allow for variations in the quantity of capital services, κK. Therefore, an expansion of L in an economic boom goes along with a reduction in Y/L, whereas a decrease of L in a recession goes along with a rise in Y/L. These results follow from the nature of the production function—for given A and κK, an increase in L results in a smaller average product of labor, Y/L. Consequently, the New Keynesian model predicts, counterfactually, that Y/L would be countercyclical.

Economists sometimes use the idea of **labor hoarding** to improve the model's predictions about labor productivity. Because of costs of hiring and firing workers, employers are motivated to retain workers during temporary downturns. Therefore, businesses may "hoard labor" in recessions as a cost-effective way of having labor available for the next upturn. Although labor input, L, still falls in a recession, it falls by less than it would if not for the hoarded labor. Moreover, during a recession, the "excess" labor may not actually produce much output. The workers may be exerting less than full effort on the job or may be performing maintenance tasks that do not show up in measured output.[8] In either case, measured output per worker, Y/L, would be relatively

[7] A procyclical pattern for the capital utilization rate, κ, reinforces the tendency for Y/L to rise in a boom and fall in a recession.
[8] Fay and Medoff (1985) found from a survey of 168 manufacturing companies that the typical firm responded to a recession by assigning an additional 5% of its work hours to maintenance, overhaul of

low in a recession. Thus, even if there are no technological shifts that affect the production function, this approach may explain why measured labor productivity is procyclical.[9]

B. Shocks to aggregate demand

Our discussion focused on the economy's responses to an increase in the quantity of money, M. However, the key part of the analysis was that each firm experienced an increase in the demand for its goods, $Y^d(j)$, while its price, $P(j)$, was held fixed. The same results apply if $Y^d(j)$ rises for other reasons, that is, if the aggregate demand for goods increases.

One way for aggregate demand to rise is for households to shift away from current saving and toward current consumption, C. That is, households become less thrifty for reasons not explained by the model. The increase in consumer demand adds to the demand, $Y^d(j)$, experienced by the typical firm.[10] In the New Keynesian model with a fixed price level, P, this expansion of demand leads, over some range, to higher real GDP, Y, and labor, L.[11]

Another possibility is that the government could boost the aggregate demand for goods by increasing its purchases of goods and services, G. This expansion would again

---

equipment, training, and other activities that do not show up in measured output. This reallocation of labor can help to explain why measured output per worker, Y/L, tends to be low during a recession.

[9] A richer version of the New Keynesian model has a different way of explaining why observed labor productivity is procyclical. The new feature is that goods produced by firms serve not only as final products but also as intermediate inputs for other firms. We already found that a monetary expansion lowers markup ratios. This reduced markup decreases the cost of intermediate inputs and thereby motivates firms to use more of these inputs. With an expansion of intermediate inputs, average labor productivity, Y/L, can rise when L increases.

[10] The government might also stimulate consumer demand by cutting taxes. If households regard themselves as wealthier when taxes are cut—unlike the Ricardian cases explored in chapter 14—current consumer demand would increase.

[11] As before, the rise in Y increases the demand for money, M^d, and this response dampens the expansion of Y.

raise the demand, $Y^d(j)$, seen by the typical firm. As before, if the price level, P, is held fixed, real GDP, Y, and labor, L, tend to rise.

The New Keynesian model has the property that an increase in the aggregate demand for goods may end up increasing real GDP, Y, by even more than the initial expansion of demand. That is, there may be a **multiplier** in the model—the rise in Y may be a multiple greater than one of the rise in demand.

The reasoning is that, at a fixed price level, P, the initial rise in aggregate demand leads to an equal size increase in production, Y. At least this response applies if all firms have markup ratios significantly above one and are, therefore, willing to meet fully the extra demand at a fixed P. The expansion of Y corresponds to increases in real income, notably in real labor income, $(w/P)\cdot L$. This added income motivates households to raise consumer demand, which provides another boost to the demand for each firm's goods, $Y^d(j)$. The further increase in production, Y, generates the multiplier.

The Keynesian multiplier is a neat theoretical result. The problem, however, is that no one has ever verified empirically the existence of this multiplier. For example, we found in chapter 12 that it was difficult to document in the U.S. data even a positive effect from changes in government purchases, G, on real GDP, Y. The positive relation was clear only for the large temporary expansions of G that applied during major wars. Even in these cases, the response of Y was less than the increase in G, that is, the multiplier was less than one.

Evidence on the stickiness of prices

A recent study by Mark Bils and Peter Klenow (2004) quantifies the extent of stickiness in prices for the goods and services contained in the U.S. consumer price index (CPI). Each month, the Bureau of Labor Statistics (BLS) collects prices on about 75,000 individual items. The flexibility of prices varies a lot depending on the type of good or service. Some types saw price changes almost every month—these include gasoline, airline fares, and fresh produce. Others changed infrequently—including vending machines, newspapers, and taxi fares. Overall, from 1995 to 2002, 22% of the individual items had changes in prices from one month to the next. The median duration of prices was 4-5 months.[12] From this evidence, we can conclude that price stickiness is significant but not of very long duration for the typical item.

Mikhail Golosov and Robert Lucas (2003) used the Bils and Klenow data to estimate how important the stickiness of prices is for U.S. economic fluctuations. In the Golosov-Lucas model, one reason that firms want to change prices is individual shocks, which affect a specific firm's product demand or technology. For example, a rise in individual product demand motivates each firm to raise its relative price—$P(j)/P$ in the model that we worked out before.

[12] Earlier studies of price stickiness include Carlton (1986), Cecchetti (1986), Kashyap (1995), and Blinder, et al (1998). These studies concluded that price stickiness was more important than that found by Bils and Klenow (2004). The reason is that the earlier studies looked only at a narrow range of products, such as newspapers and items listed in catalogs, which have above average price stickiness.

The second reason to change prices is an economy-wide monetary disturbance. For example, a monetary expansion motivates each firm to raise its nominal price, *P(j)*.

Both types of price changes incur a menu cost. Because of these costs of changing prices, an individual firm does not always adjust its price, *P(j)*, in response to an individual shock or a monetary disturbance. Suppose that *[P(j)]** represents a firm's "ideal price," that is, the price that would be chosen if menu costs were zero. Each firm will find it optimal to make a price change when *P(j)* deviates substantially from *[P(j)]**. When a price change occurs, *P(j)* will typically adjust by a substantial amount, perhaps 5-10% upward or downward.

Golosov and Lucas construct a model that has an array of individual firms, each of which has the same menu cost for changing its price. (An extension to allow for differences in menu costs turns out not to affect the main results.) A higher menu cost motivates less frequent price changes. Golosov and Lucas assume that the menu cost takes on a value that generates the frequency of price change found in the Bils-Klenow data. Thus, in the Golosov-Lucas model of the U.S. economy, individual firms adjust their prices on average every 4-5 months.

One finding in the model is that, for the U.S. economy, individual shocks are responsible for most of the price changes. That is, most prices change because of shifts to individual demand or technology, rather than economy-wide monetary shocks. The situation is different for an economy with high and variable inflation—Golosov and Lucas study Israel in the late 1970s and early 1980s. In that environment, most price changes occur because of economy-wide monetary shocks.

In the Golosov-Lucas model, monetary shocks affect labor input and output, as in the New Keynesian model that we studied. However, it turns out for the U.S. economy that observed monetary fluctuations account for only a small fraction of the observed fluctuations in real GDP. Golosov and Lucas conclude that, although money is non-neutral, monetary shocks play a minor role in economic fluctuations.

II. Money and Nominal Interest Rates

In practice, central banks such as the Federal Reserve tend to express monetary policy in terms of short-term nominal interest rates, rather than monetary aggregates. In the United States, especially since the early 1980s, the Fed focuses on the Federal Funds rate—the overnight nominal interest rate on funds exchanged between financial institutions, such as commercial banks.

The Federal Reserve's Federal Open Market Committee or FOMC meets eight or more times a year. At each meeting, the FOMC adopts a target for the Federal Funds rate. Then the FOMC instructs the Federal Reserve's trading desk to conduct open-market operations to achieve the desired target for the funds rate. Open-market operations are exchanges between the monetary base (currency plus reserves held by depository institutions at the Fed) and interest-bearing assets, principally U.S. Treasury securities. In an expansionary operation, the Fed creates new monetary base to buy interest-bearing assets. In a contractionary operation, the Fed sells interest-bearing assets from its portfolio and uses the proceeds to reduce the monetary base.

We can think of the relation between money and nominal interest rates from our familiar equilibrium condition (equation [10.3] in chapter 10) that the nominal quantity of money, M, equal the nominal quantity demanded, $P \cdot £(Y, i)$:

$$M = P \cdot £(Y, i). \tag{16.8}$$

In our model, we have thought of M as nominal currency. Now we should add to currency the nominal reserves held by depository institutions at the Fed. These reserves, like paper currency, are non-interest-bearing liabilities of the Federal Reserve. The sum of currency (including amounts held in vaults of depository institutions) and reserves held at the Fed is the monetary base. The monetary base is the monetary aggregate affected directly by the Fed's open-market operations. Through these operations, the Fed can control the quantity of monetary base on a day-by-day basis.

The equilibrium condition in equation (16.8) specifies a relationship between the nominal monetary base, M, and the determinants of the nominal demand for money: the price level, P, real GDP, Y, and the nominal interest rate, i. In the New Keynesian model that we have been using, P is fixed in the short run. Thus, in the short run, if M increases, equilibrium requires some combination of higher Y or lower i to raise the nominal quantity of money demanded by the same amount as the increase in M. For a given Y, equation (16.8) says that a higher M has to match up with a lower i.

In our previous analysis, we thought of an expansionary monetary shock as an increase in M. Then we argued that, if firm j had with a fixed price, $P(j)$, the monetary change would raise the demand for firm j's goods, $Y^d(j)$. Now we can think instead of an expansionary monetary disturbance as a decrease in the nominal interest rate, i. Again, the key point is that this monetary stimulus raises the demand for goods, $Y^d(j)$.

If equation (16.8) were a fixed relationship, the central bank could operate equivalently by changing the monetary base, M, or the nominal interest rate, i. However, in practice, the real demand for money, $£(Y, i)$, tends to fluctuate a good deal. As an example, we noted in chapter 10 that this demand has a lot of seasonal variation. The Fed accommodates these movements in real money demand by introducing an appropriate amount of seasonal variation in M. This seasonal in the monetary base avoids having a seasonal in i—in fact, the elimination of seasonal variations in nominal interest rates was one of the reasons for the creation of the Federal Reserve in 1914. For discussions, see Truman Clark (1986) and Jeffrey Miron (1986).

More generally, by targeting the nominal interest rate, i, the Fed is led automatically to carry out the volume of open-market operations needed to get the changes in the monetary base, M, that balance changes in the nominal quantity of money demanded, $P\cdot £(Y, i)$. That is, in equation (16.8), if M and $£(Y, i)$ each change by the same amount, no changes have to occur in i (or P or Y). It would be impossible to achieve the same result by designating in advance the precise time path for the monetary base or some other monetary aggregate. Such a designation would require knowledge about future levels of real money demand, $\pounds(Y, i)$.[13] Because this knowledge is unattainable, central banks tend not to conduct monetary policy by specifying the future path of a nominal monetary aggregate. More specifically, central banks have rejected proposals, originally put forward by Milton Friedman (1959, pp. 90-93), to have a designated monetary aggregate grow at a constant rate. Such constant-growth-rate rules for money require i (or P or Y) to respond to each variation in the real demand for money.

Since the early 1980s, the Federal Reserve has adjusted its target for the Federal Funds rate, i, to achieve a moderate and stable rate of increase in the price level, P. That is, the Fed has generated a moderate and stable inflation rate, π. In recent years, the Fed seems to regard an inflation rate of around 2% as acceptable, so that π above 2% is viewed as too high and π below 2% as too low. If π is greater than 2% for some time, the Fed tends to take contractionary action—raising i by means of contractionary open-market operations. Conversely, if π is below 2% for some time, the Fed tends to take expansionary action—lowering i by means of expansionary open-market operations.

Figure 16.2 shows the history of the Federal Funds rate and the inflation rate (based on the GDP deflator) from 1954.1 to 2004.3. One important observation is that the Fed seems to have changed the way it sets the Funds rate in response to inflation. For example, from 1972 to 1975, when the inflation rate rose sharply, the Funds rate increased only along with the inflation rate. Then, when the inflation rate came down after 1975, the Funds rate declined along with it. In contrast, in 1979-80, the rise in the inflation rate resulted in an even greater increase in the Funds rate, which soared to 20%

[13] Similarly, in a seasonal context, the Fed does not know in advance the precise relation of the real demand for money to the calendar date. By targeting the nominal interest rate, i, the Fed is led to the seasonal variation in the monetary base, M, that avoids a seasonal in i. This procedure works even when changes occur in the seasonal pattern of real money demand.

in 1981. The Fed brought the Funds rate down below 8% only after the inflation rate fell sharply in the early and mid 1980s.

The change in the Fed's monetary policy, whereby the Funds rate reacts strongly to inflation, was initiated by Paul Volcker, who was chair of the Fed from August 1979 until August 1987. His procedures have been continued by Alan Greenspan, who became chair in August 1987. The new policy seems to have succeeded in maintaining low and stable inflation. As Figure 16.2 shows, the inflation rate since 1985 averaged a little over 2% and has fluctuated within a narrow range. Similar successes in curbing inflation has been achieved since the 1980s by the central banks of other advanced countries.

Many of the movements in the Federal Funds rate during the Volcker and Greenspan years were reactions to economic variables other than inflation. The Fed tended to raise the Funds rate when the real economy was strong, gauged especially by high employment growth and a low unemployment rate. An example is the steep rise in the Funds rate from 3% at the end of 1993 to 6% in mid 1995, apparently in response to strength in the labor market. Conversely, the Fed reduced the Funds rate when the real economy was weak. An example is the cut in the Funds rate from over 6% in mid 2000 to the remarkably low level of 1% in 2003, apparently in reaction to the weak labor market.

At this point, economists are confident that the response of monetary policy to inflation has been beneficial. That is, the success in curbing inflation seems to derive from the policy of raising nominal interest rates sharply when inflation rises and cutting rates sharply when inflation falls. This policy can be expressed alternatively has having contractionary open-market operations when inflation increases and expansionary open-market operations when inflation decreases.

The benefits from the other parts of the Fed's monetary policy are less clear. That is, we do not know whether the economy has performed better or worse because the Fed has adjusted nominal interest rates in response to strength or weakness in the real economy

III. The Keynesian Model—Sticky Nominal Wage Rates

As mentioned before, the model in Keynes's (1936) *General Theory* relies on sticky nominal wage rates. We can bring out the essence of this model by assuming that prices of goods are perfectly flexible. In fact, we can return to the setting from previous chapters in which the suppliers and demanders of goods are perfect competitors. In this setting, the single nominal price, P, prevails in the market for goods.

Keynes just assumed that the nominal wage rate, w, was sticky. That is, he assumed that w did not adjust rapidly to clear the labor market. Moreover, Keynes focused on a case in which w was above its market-clearing level. Consequently, when the price level, P, adjusts to balance the quantities of goods supplied and demanded, the real wage rate, w/P, will end up higher than its market-clearing value.

In the Keynesian model, the labor market looks as shown in Figure 16.3. In this graph, the labor demand and supply curves are the same as those in our previous model (see Figure 8.3 in chapter 8). Note that an increase in the real wage rate, w/P, lowers the quantity of labor demanded, L^d, but raises the quantity supplied, L^s. The only difference from before is that the nominal wage rate, w, is assumed not to adjust to generate the real wage rate, $(w/P)^*$, that balances the quantities of labor demanded and supplied. Instead, the prevailing real wage rate, shown as $(w/P)_1$, is assumed to be above $(w/P)^*$.

At the real wage rate $(w/P)_1$ shown in Figure 16.3, the quantity of labor supplied is greater than the quantity demanded. Since the quantities supplied and demanded are unequal, we have to consider how the quantity of labor, L, is determined. Our assumption is that L equals the smaller of the amounts demanded and supplied—in this case, the quantity demanded, L^d. Labor input cannot be higher than this amount, because some demander of labor would then be forced to employ more labor than the quantity desired at the given real wage rate, $(w/P)_1$. In other words, we assume that the labor market respects the principle of **voluntary exchange**. No market participant can be forced to hire more labor than the amount desired—or to work more than the amount desired—at the prevailing real wage rate.

Notice in Figure 16.3 that the quantity of labor supplied, L^s, at the given real wage rate, $(w/P)_1$, is greater than the quantity demanded, L^d, which equals the quantity of labor, L. The usual assumption about markets is that the nominal wage rate, w, would decline in this situation. That is, the eager suppliers of labor would compete for jobs by bidding down w. However, this response is ruled out by assumption in the Keynesian model, at least in the short run.

When the quantity of labor supplied, L^s, is greater than the quantity of labor, L, the model does not tell us how the available labor is divided up among the households. We have to assume that some rationing process allocates labor—employment or work hours—among those seeking work. In the Keynesian model, the excess of the quantity of labor supplied (at the given real wage rate, $(w/P)_1$) over L is called **involuntary**

unemployment. This amount is shown in green in Figure 16.3. Involuntary unemployment is the difference between the quantity of labor that households would like at $(w/P)_1$—that is, the quantity supplied, L^s—and the quantity they actually get, L.

Suppose now that a monetary expansion raises the price level, P, from P_1 to P_2. If the nominal wage rate, w, does not change, the rise in P lowers the real wage rate, w/P. We assume in Figure 16.4 that the increase in P from P_1 to P_2 reduces the real wage rate from $(w/P)_1$ to $(w/P)_2$. This fall in w/P raises the quantity of labor demanded, L^d, and, therefore, increases labor input from L_1 to L_2. Hence, in the model with sticky nominal wage rates, a monetary expansion raises labor, L. The increase in L leads through the production function to an expansion of real GDP, Y.

As long as the nominal wage rate, w, is fixed, monetary expansions reduce the real wage rate, w/P, and raise labor, L, through the mechanism shown in Figure 16.4. This process can continue until w/P falls to its market-clearing value, $(w/P)^*$. At that point, labor reaches its market-clearing level, L^*.[14]

We see from Figure 16.4 that a monetary expansion raises the quantity of labor, L, and thereby real GDP, Y, by lowering the real wage rate, w/P. The Keynesian model is similar to the New Keynesian model in predicting that nominal money, M, and labor, L, would be procyclical. However, unlike the New Keynesian model, the Keynesian model predicts that w/P would be countercyclical—low when L and Y were high and high when L and Y were low. We have stressed that w/P typically moves in a procyclical manner.

---

[14] Further monetary expansion would lower w/P below $(w/P)^*$. In this situation, the quantity of labor demanded, L^d, becomes larger than the quantity supplied, L^s. According to the principle of voluntary exchange, the quantity of labor, L, would then equal the quantity supplied, L^s. In this circumstance, reductions in w/P brought about by further monetary expansions would lower the quantity of labor supplied, L^s, and, thereby, reduce L. Thus, in the Keynesian model, too much monetary expansion would have adverse consequences.

Therefore, the Keynesian model has difficulty in explaining the observed cyclical behavior of *w/P*. In this respect, the model has the same flaw as the price-misperceptions model. We noted in chapter 15 that the price-misperceptions model also predicts—counterfactually—a countercyclical pattern for *w/P*.

Keynes himself recognized that his model had problems with respect to its predictions about the cyclical behavior of the real wage rate, *w/P*. However, he did not come up with convincing solutions for this problem. One of the main motivations for the New Keynesian model—developed earlier in this chapter—is that it eliminates the counterfactual predictions for *w/P*. That is, *w/P* is procyclical in the New Keynesian model. On the other hand, the New Keynesian model relies on sticky nominal prices, and many economists believe—like Keynes—that sticky nominal wage rates are more important in practice.

Finally, the Keynesian model is similar to the New Keynesian model with respect to predictions about average labor productivity, *Y/L*. In both models, the key assumption is that the technology level, *A*, is fixed. Therefore—at least for a fixed quantity of capital services, κK—an expansion of *L* goes along with a reduction in *Y/L,* whereas a decrease of *L* goes along with a rise in *Y/L*. This result follows from the property of diminishing average product of labor. Consequently, the Keynesian model predicts, counterfactually, that *Y/L* would be countercyclical. As with the New Keynesian model, the concept of labor hoarding has been offered to explain this failing of the Keynesian model.

The real wage rate during the New Deal

Typically, the real wage rate, w/P, is procyclical—high when the economy is doing well and low when the economy is doing badly. However, Harold Cole and Lee Ohanian (2004) find that w/P behaved in an unusual manner during President Franklin Roosevelt's New Deal period from 1934 to 1939. At this time, the economy was recovering from the Great Depression, which was at its worst in 1933. According to Figure 7.4 in chapter 7, real GDP was 17% below trend in 1933. Thereafter, the economy improved but at a pace that many economists regard as surprisingly slow. For example, real GDP was still 9% below trend in 1939 (following the recession of 1938).

Cole and Ohanian attribute much of the slow economic recovery from 1934 to 1939 to New Deal policies that limited competition in product markets and increased labor bargaining power. Notably, these policies raised nominal and real wage rates in manufacturing and other sectors covered by the National Industrial Recovery Act (in place for 1933-35 until it was declared unconstitutional) and the National Labor Relations Act (in effect from 1935 onward). For example, in manufacturing, the real wage rate, w/P, in 1939 was 20% higher, relative to trend, than in 1933.[15] Cole and Ohanian argue that the government's policies for promoting an artificially high w/P and for restricting competition among businesses were harmful to the economy. Specifically, they

[15] See Cole and Ohanian (2004, Table 2).

estimate that the rate of economic growth from 1934 to 1939 would have been substantially higher if the New Deal policies had not been in place. Interestingly, Roosevelt came to the same conclusion in 1939, when he abandoned his efforts to limit competition in the economy.

IV. Long-Term Contracts and Sticky Nominal Wage Rates (optional material)

For many workers, nominal wage rates are set for one or more years by the terms of agreements made with employers. These agreements are sometimes formal contracts between firms and labor unions. More commonly, firms and workers have implicit contracts that specify in advance the nominal wage rate over some period, often a fiscal or calendar year. The existence of these agreements has been offered as a defense for the Keynesian assumption that the nominal wage rate, w, is sticky. The contracting approach has also been used to explain why some nominal prices are sticky, for example, the prices of intermediate goods sold to businesses by regular suppliers.

There are many good reasons why trading partners would specify in advance the wage rates or prices on which services and goods will be exchanged. One point is that this presetting of wage rates or prices can prevent one party from demanding "unreasonable" terms, ex post. That is, a wage or price agreement can alleviate what economists call the "hold-up problem." In the absence of a contract, an employer might, for example, lower the nominal wage rate after an employee had incurred substantial costs in moving to a job. Similarly, a supplier might raise the price for materials at a time when delays in a construction project have become prohibitively expensive. Presetting the wage rate or price can avoid some of these problems.

Suppose that an employer and employee agree on a fixed nominal wage rate, w, for the next year.[16] A natural choice is to set w equal to the best estimate of the average market-clearing nominal wage rate, w^*, that will prevail during the year. Although the chosen w may be a rational expectation of w^*, unanticipated events lead to mistakes. For example, if the inflation rate, π, is surprisingly high, the average price level, P, during the year will be higher than anticipated. If nothing else changes in the economy, the market-clearing nominal wage rate, w^*, would rise along with P to maintain the market-clearing real wage rate, $(w/P)^*$. In this case, the fixed nominal wage rate, w, falls short of the average w^* for the year. Conversely, if π is surprisingly low, the fixed w would be greater than the average w^* for the year.

When the contract expires, the employer and employee agree on a new nominal wage rate, w, for the next year. This new w takes account of events during the current year, including the inflation rate, π. Thus, if expectations are rational, mistakes in the setting of w for this year—due perhaps to underestimation of inflation—tend not to be repeated the next year. The rationality of expectations implies also that deviations between w and w^* would not be systematically greater than or less than zero. Thus, the contracting perspective does not support the Keynesian emphasis on situations in which w is greater than w^*.

---

[16] The contracting approach motivates the presetting of a real wage rate, w/P, rather than a nominal wage rate, w. Yet most labor agreements in advanced economies are not explicitly "indexed," that is, do not contain automatic adjustments of nominal wage rates for changes in the price level, P. Apparently, firms and workers find it convenient to specify agreements in terms of the standard nominal unit of account—such as the U.S. dollar—even though the future price level is uncertain. There is evidence, however, that higher and more variable inflation tends to generate shorter contracts and more frequent use of formal indexation of nominal wage rates.

At any point in time, the economy has an array of existing labor contracts, each of which specifies a nominal wage rate, w, that likely deviates somewhat from the market-clearing value, w^*. Some of these agreements have w greater than w^* and others have w less than w^*. However, aggregate shocks can create differences between the economy-wide averages of w and w^*. For example, unexpectedly low inflation tends to make w greater than w^* throughout the economy. Some macroeconomists have used this result to explain why a monetary contraction reduces employment and output. That is, the contracting approach has been used to fill an important gap in the Keynesian model—explaining why the nominal wage rate, w, is sticky.[17]

Unfortunately, this application of the contracting approach encounters logical problems. The Keynesian results emerge when the nominal wage rate, w, is fixed at too high a level **and** when the principle of voluntary exchange determines the quantity of labor, L. For example, when w and, hence, w/P are too high, so that the quantity of labor supplied, L^s, is greater than the quantity demanded, L^d, voluntary exchange dictates that L equal the quantity demanded. This approach makes sense for an impersonal market. However, the idea is typically not sensible in the context of a long-term contract, which is supposed to be the rationale for sticky nominal wage rates.

In an enduring relationship, in which explicit or implicit contracts arise, the trading parties do not have to change prices or wage rates at every instant to get the "right" behavior of quantities. Workers can, for example, agree in advance that they will work harder when there is more work to do—that is, when the demand for a firm's product is high—and work less hard when there is little work. Unlike in an impersonal

[17] The original applications of the contracting approach to macroeconomics were by Gordon (1974), Azariadis (1975), and Baily (1974). Applications to monetary shocks were made by Gray (1976), Fischer (1977), and Taylor (1980).

auction market, these efficient adjustments in work and production can occur even if wage rates do not change from day to day.[18] The important point is that, in the context of labor contracts, stickiness of nominal wage rates does not necessarily cause errors in the determination of labor and production.

To take a concrete example, suppose that inflation is sometimes lower than expected and sometimes higher. Rational firms and workers know that inflation—if not accompanied by real changes in the economy—does not alter the efficient levels of labor and production. Therefore, it makes sense to agree on a contract that insulates the choices of labor and output from the inflation rate. Over many years, when the effects of unanticipated inflation on real wage rates tend to average out, both parties to the contract would benefit from this provision. However, if inflation is high and unpredictable, firms and workers prefer either to index nominal wage rates to the price level or to renegotiate contracts more frequently.

An important lesson from the contracting approach is that stickiness of the nominal wage rate, w, need not lead to the unemployment and underproduction that appears in the Keynesian model. Within a long-term agreement, it is unnecessary for w to move all the time in order for the economy to attain the market-clearing quantity of labor, L^*. Thus, instead of supporting the Keynesian perspective, the contracting analysis demonstrates that observed stickiness of nominal wage rates (and prices) may not matter very much for the workings of the macro economy.

[18] However, for large short-term increases in labor, contracts often prescribe overtime premiums or other types of bonuses.

Empirical evidence on the contracting approach

A number of empirical studies provide evidence about the macroeconomic implications of the contracting approach. Shaghil Ahmed (1987a) used a data set for 19 industries in Canada over the period 1961-74. He used these data because an earlier study by David Card (1980) calculated the amount of indexation—automatic adjustment of nominal wages rates for changes in the price level—in each industry's labor contracts. Indexation ranged from zero to nearly 100%. According to theories in which labor contracts are the basis for the Keynesian model, industries with little indexation should show substantial responses of real wage rates and, hence, of employment and output, to nominal shocks. Industries with lots of indexation would be affected little by nominal disturbances.

Ahmed found that monetary shocks had positive effects on hours worked in most of the 19 industries. The important point for present purposes, however, is that the extent of an industry's response to these shocks bore no relation to the amount of indexation in the industry. Those with lots of indexation were as likely as those with little indexation to respond to monetary shocks. This finding is damaging to theories that use long-term contracts as the basis for the Keynesian sticky-wage model.

Mark Bils (1989) studied labor contracts for 12 manufacturing industries in the United States. He reasoned that, if the signing of new contracts was important, we should observe unusual behavior of employment and real wage rates just after these signings. His results were mixed. Some industries, notably motor vehicles, showed substantial changes in employment just after the implementation of new labor agreements. Prior changes in employment tended to be reversed just after a new contract came into effect. These results, although applying only to a few industries, support the contracting approach. However, Bils did not find any corresponding changes in real wage rates after new labor contracts were implemented. Since these changes in real wage rates are central to the contracting approach, it is difficult to reconcile this part of Bils's findings with that approach.

A recent study by Giovanni Olivei and Silvana Tenreyro (2004) suggests that the contracting approach may be important for understanding the real effects of monetary policy. They first observe that a preponderance of firms set wage rates toward the end of each calendar year, with the changes taking effect in January of the next year. In the contracting approach, this timing means that monetary disturbances that occur toward the end of the calendar year would be undone within a few months by the changes of wage rates in the next annual adjustment. In contrast, monetary disturbances at the beginning of the calendar year would take up to 12 months to be undone by the next adjustment.

Using this conceptual framework, Olivei and Tenreyro investigate whether the response of real GDP to monetary shocks looks different depending on the quarter of the year in which the shock occurs. They measure the shocks by unusual movements in the Federal Funds rate, which is the short-term nominal interest rate that the Fed monitors closely. The unusual movements are those that cannot be explained by prior variations in real GDP, the GDP deflator, and commodity prices. The main finding, for the period 1966 to 2002, is that the response of real GDP to a Funds-rate shock is substantial when the shock takes place in the first or second quarter of the year. A decline in the interest rate by one-quarter of a percentage point is estimated to raise real GDP by an average of 0.2% over the following two years. However, the response is smaller if the shock to the Funds rate occurs in the third or fourth quarter of the year. In this case, a decrease in the funds rate by one-quarter of a percentage point is estimated to raise real GDP by an average of less than 0.1% over the following two years. Olivei and Tenreyro suggest that the difference in response arises because nominal wage rates tend to be sticky during the calendar year but flexible from the end of one calendar year to the beginning of the next.

V. Summing Up

Another possible explanation for monetary non-neutrality involves stickiness of nominal prices and wage rates. Generally, this stickiness reflects costs of changing prices and wages.

The New Keynesian model features stickiness in prices of goods. Individual firms with monopoly power set prices as markups over nominal marginal costs of production. If prices are fixed in the short run, firms meet expansions of demand—over some range—by raising production and labor input. Therefore, a monetary expansion increases the economy-wide quantity of labor demanded. The boost to labor demand raises the real wage rate and, if the labor-supply curve slopes upward, the quantity of labor input. This expansion of labor allows for an increase in real GDP. The model thereby predicts that nominal money, labor, and the real wage rate will be procyclical. The prediction for procyclical real wage rates—which is consistent with the data—distinguishes this model from the price-misperceptions model. However, the New Keynesian model has the counterfactual prediction that labor productivity will be countercyclical. The idea of labor hoarding has been offered to explain this shortcoming.

The older style Keynesian model relied on sticky nominal wage rates. If the nominal wage rate is typically too high, the quantity of labor supplied tends to exceed the quantity demanded. Employment equals the quantity demanded, and the shortfall from the quantity supplied equals the amount of involuntary unemployment. In this setting, a monetary expansion lowers the real wage rate and, thereby, raises the quantity of labor demanded and, hence, the level of employment. However, this model predicts, counterfactually, that real wage rates will be countercyclical.

Questions and Problems

Mainly for review

16.1. What is involuntary unemployment?

16.2. Explain how an increase in the quantity of money, M, reduces the nominal interest rate, i, in the new Keynesian model. Why does this effect not arise in the market-clearing model?

Problems for discussion

16.x. The paradox of thrift

Suppose that households become thriftier in the sense that they decide to raise current saving and reduce current consumer demand.

a. In the New Keynesian model, what happens to real GDP, Y, and labor, L?

b. What happens to the amount of saving? If it decreases, there is said to be a paradox of thrift.

c. Can there be a paradox of thrift in the market-clearing model considered in chapters 7 and 8?

16.x. The Keynesian multiplier

a. Explain why there can be a multiplier in the new Keynesian model.

How is the size of the multiplier affected by the following:

b. adjustments of the price level, P?

c. the extent to which markup ratios exceed one?

d. reactions of money demand, M^d, to real GDP, Y?

16.x. Perceived wealth in the New Keynesian model

Suppose that the U.S. President makes a speech and announces that we are all wealthier than we thought. If we all believe the President, what does the New Keynesian model predict for changes in real GDP, Y, and labor, L? Do we actually end up being "wealthier?" Contrast these predictions with those from the market-clearing model in chapters 7 and 8.

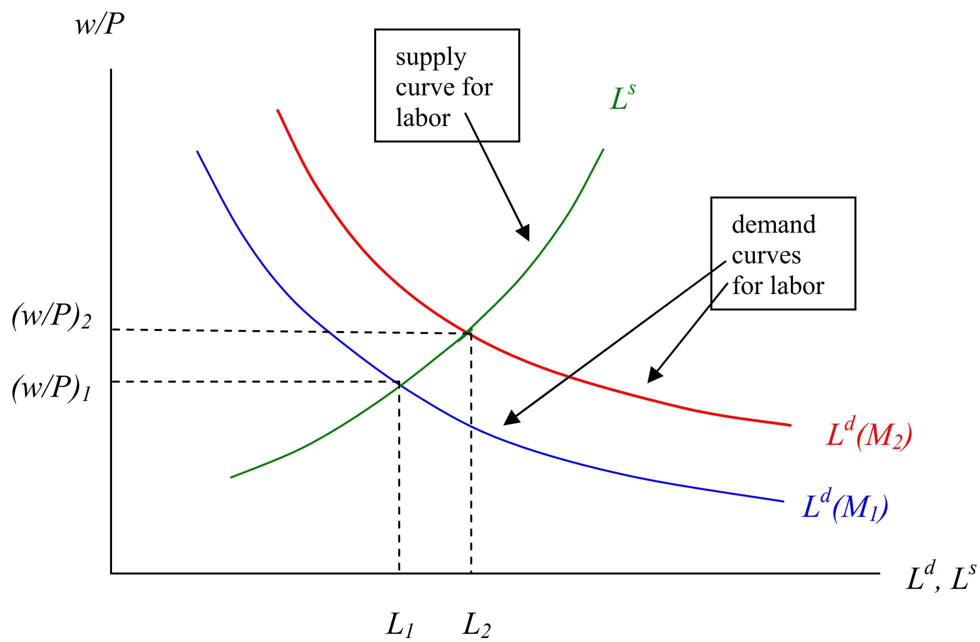


Figure 16.1

Effect of a Monetary Expansion in the New Keynesian Model

When the nominal quantity of money is M_1, the demand for labor, labeled $L^d(M_1)$, slopes downward versus the real wage rate, *w/P*, and is given by the blue curve. When the nominal quantity of money rises to M_2—but the price level, *P*, is held fixed—the demand for labor, labeled $L^d(M_2)$ along the red curve, is larger at any given *w/P*. The supply of labor, L^s, in green, slopes upward versus *w/P*. The increase in the quantity of money from M_1 to M_2 raises the real wage rate from $(w/P)_1$ to $(w/P)_2$ and increases labor input from L_1 to L_2.

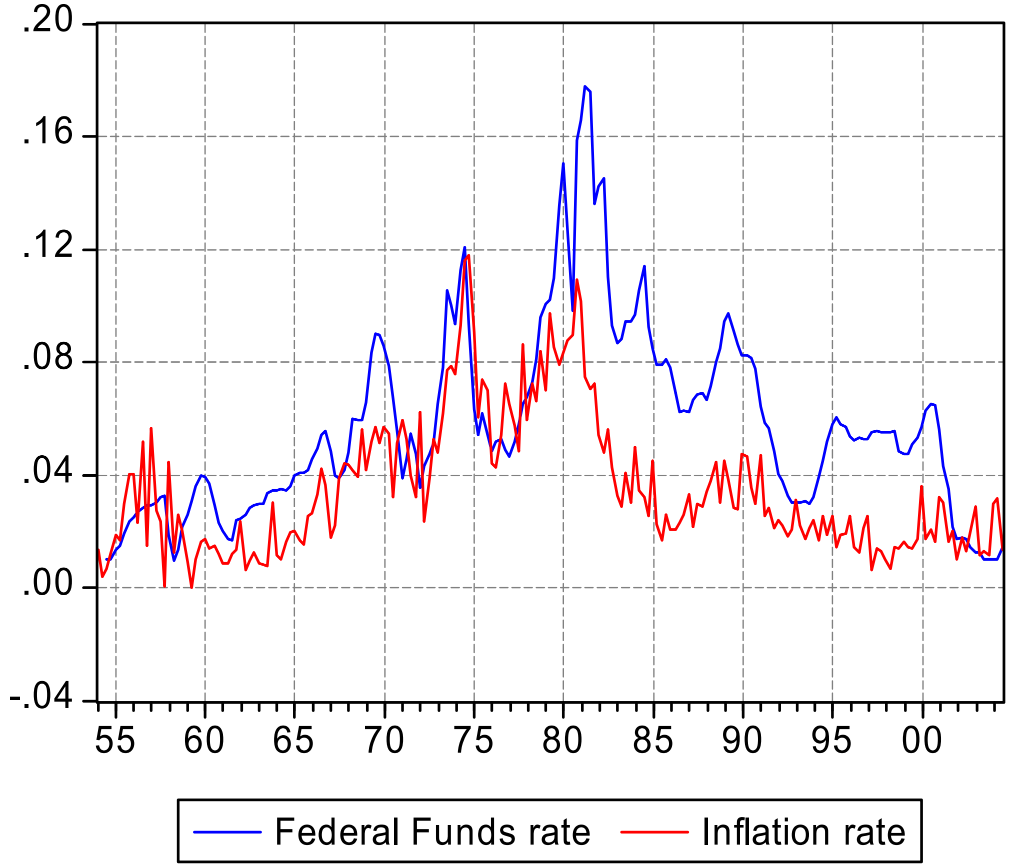


Figure 16.2

The Federal Funds Rate and the Inflation Rate

The blue line is the Federal Funds rate, the short-term nominal interest rate controlled by the Federal Reserve. The red line is the inflation rate, based on the GDP deflator.

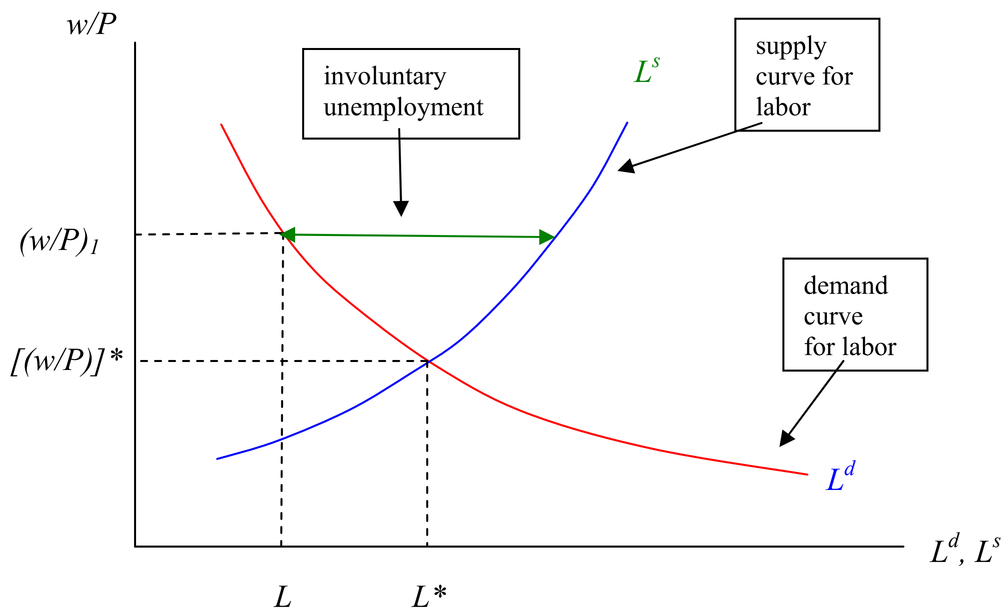


Figure 16.3

The Labor Market in the Keynesian Model with Sticky Nominal Wage Rates

In the Keynesian model, the nominal wage rate, w, is fixed above its market-clearing value, w^*. When the price level is P_1, the real wage rate, $(w/P)_1$, is greater than the market-clearing value, $(w/P)^*$. At $(w/P)_1$, the quantity of labor supplied, along the L^s curve, exceeds the quantity demanded, along the L^d curve. The quantity of labor, L, equals the quantity demanded and is less than the market-clearing value, L^*.

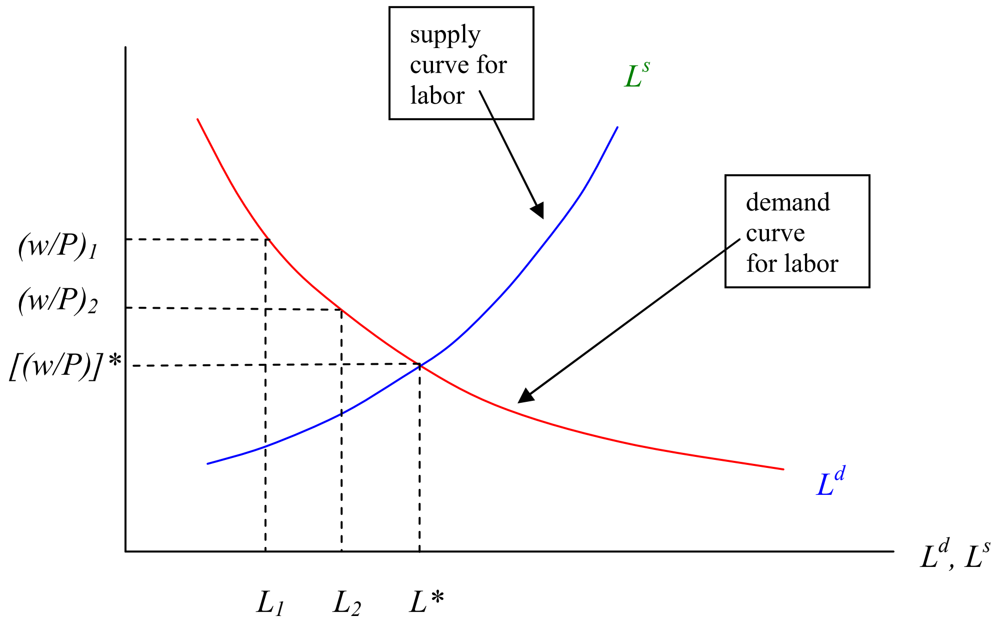


Figure 16.4

Effect of a Monetary Expansion in the Keynesian Model with Sticky Nominal Wage Rates

In the Keynesian model, the nominal wage rate, w, is fixed above its market-clearing value, w^*. When the price level is P_1, the real wage rate is $(w/P)_1$, which is greater than the market-clearing value, $(w/P)^*$. Labor, L_1, equals the quantity of labor demanded at $(w/P)_1$. A monetary expansion raises the price level to P_2 and, thereby, lowers the real wage rate to $(w/P)_2$. Labor rises to L_2, the quantity of labor demanded at $(w/P)_2$. A sufficient monetary expansion could raise L as high as the market-clearing value, L^*.

January 12, 2005

Chapter 17

World Markets in Goods and Credit

Thus far, we carried out our analysis for a single, closed economy. Therefore, we neglected the interactions among countries on international markets. Many macroeconomists, especially those in the United States, focus on a closed-economy framework. One justification is that the U.S. economy represents a large share of the world economy, which really is a closed economy (if we neglect trade with Mars). In addition, compared to many of the world's small economies, a minor fraction of U.S. production and expenditure involves international trade. Another rationale for the neglect of world markets, applicable particularly to the 1950s and 1960s, was that various restrictions inhibited the flows of goods and credit from one country to another.

With the opening up of international markets—that is, **globalization**—over the last fifty years, the practice of ignoring the rest of the world became increasingly unsatisfactory even for the U.S. economy. The ratio of U.S. imports to GDP rose from 4% in 1950 to 15% in 2004, while the ratio of exports to GDP increased from 4% to 10%. Recent years have also featured increased borrowing from foreigners by the United States.

To study international trade, we have to extend our model to the world economy, which comprises many countries. We carry out the analysis from the perspective of a **home country**. We think of the rest of the world as a single entity, which we call the foreign country. Residents of the home country buy goods and services from foreigners (imports) and sell goods and services to foreigners (exports). Residents of the home country also borrow from and lend to foreigners.

Sometimes we make the assumption that the home country has a negligible effect on the equilibrium in the foreign country, that is, in the rest of the world. This assumption is satisfactory if the home country is a minor part of the world economy. The United States is an intermediate case between a small open economy and the world economy (which is a closed economy). That is, the United States is large enough to have a noticeable effect on the equilibrium of the world economy.

To extend our model to an international setting, we begin with a number of unrealistic assumptions, which we can later relax. Assume first that the goods produced in each country are physically identical. In addition, suppose that transport costs and barriers to trade across national borders are small enough to neglect. Finally, pretend at this stage that, instead of using their own money, all countries use a **common currency**, such as the U.S. dollar or the euro. The residents of each country hold money in the form of, say, U.S. dollars and quote prices in units of U.S. dollars.

Given these assumptions, goods in all countries must sell at the same dollar price, *P*. If prices differed, households would want to buy all goods at the lowest price and sell all goods at the highest price. Thus, in an equilibrium, where goods are bought and sold in all locations, all prices must be the same. This result is the simplest version of the **law of one price**. At this point, we also ignore inflation, so that the price level, *P*, in all countries is constant over time.

Suppose that each country has a central bank and that this bank holds a quantity of **international currency**. This currency could be pieces of paper denominated in U.S. dollars or other units, such as euros. The currency could also take the form of a commodity, such as gold. The precise form of the international currency is unimportant for our purposes, except that we assume that the nominal interest rate on this currency is zero.

Let $\overline{H}$ denote the world quantity of international currency, denominated in U.S. dollars. (An overbar means that the variable applies to the whole world.) We assume, for simplicity, that $\overline{H}$ is constant over time. In year *t*, the home country's central bank holds the real quantity of international currency H_t/P to facilitate transactions between domestic residents and foreigners.

Suppose that the nominal interest rate in the home country is i_t. Since we neglect inflation, the real interest rate, r_t, equals i_t. Suppose that the nominal interest rate in the rest of the world is i_t^f. Since we also neglect inflation in the rest of the world, the foreign real rate, r_t^f, equals i_t^f. In our previous analysis, we neglected differences among borrowers in creditworthiness. Now we go further to neglect differences in creditworthiness between domestic residents and foreigners. Furthermore, we assume that no transaction costs exist for carrying out financial exchanges across international borders. Given all these assumptions, the world credit markets would have to function effectively as a single market with a single nominal interest rate. That is, we have

$$i_t = i_t^f$$

and

$$r_t = r_t^f .$$

The nominal and real interest rates are the same for lenders and borrowers in all countries.

I. The Home Country and the Rest of the World

Consider the residents of the home country. We still let Y represent the real gross domestic product (GDP), which is the quantity of goods and services produced domestically. For the residents of the home country, the total of funds lent need no longer equal the total borrowed. Rather, the total amount lent on net from the home country to the rest of the world equals the total amount borrowed by foreigners from the home country. To see how this works out in terms of budget constraints, we need some new notation.

Let B_t^f be the net nominal holding of foreign interest-bearing assets by domestic residents at the end of year t. We think of these assets as bonds, but they could also be ownership rights in capital located abroad (called **foreign direct investment**). For simplicity, we assume that the domestic government neither holds foreign bonds nor borrows from foreigners. In this case, B_t^f represents the net holding of foreign bonds by domestic households.

In chapter 12, we worked out a form of the household budget constraint in equation (12.4). We can use that result to write the household budget constraint in real terms as[1]

(17.1) $C_t + (B_t – B_{t-1})/P + (M_t - M_{t-1})/P + K_t – K_{t-1} = (w/P)_t \cdot L_t + r \cdot (B_{t-1}/P + K_{t-1}) + (V_t – T_t)/P.$

[1] In equation (12.4), we neglected changes in money holdings, that is, we assumed $\Delta M_t = M_t – M_{t-1} = 0$. Equation (17.1) allows for $\Delta M_t \neq 0$.

We are assuming that the price level, P, and the real interest rate, r, are not changing over time. Note that the left-hand side has uses of households' funds: consumption, C_t, increases in real holdings of bonds and money, $(B_t–B_{t-1})/P + (M_t-M_{t-1})/P$, and increases in ownership of capital, $K_t–K_{t-1}$. The right-hand side has sources of funds: real labor income, $(w/P)_t\cdot L_t$, real income on assets, $r\cdot(B_{t-1}/P + K_{t-1})$, and real transfers net of real taxes, $(V_t–T_t)/P$.

Recall that real GDP, Y_t, equals the sum of real labor income and real rental payments:

$$Y_t = (w/P)_t\cdot L_t + (R/P)\cdot K_{t-1}$$

$$Y_t = (w/P)_t\cdot L_t + (r+\delta)\cdot K_{t-1},$$

where we used the condition that the real rental price, R/P, equals the real interest rate, r, plus the depreciation rate, δ. If we use the last expression for Y_t, we can substitute out

$$(w/P)_t\cdot L_t + r\cdot K_{t-1} = Y_t – \delta K_{t-1}$$

on the right-hand side of equation (17.1) to get

$$C_t + (B_t–B_{t-1})/P + (M_t-M_{t-1})/P + K_t–K_{t-1} = Y_t + r\cdot(B_{t-1}/P) – \delta K_{t-1} + (V_t–T_t)/P.$$

For the aggregate of households, the total of bonds held, B_t, equals the sum of domestic government bonds, B_t^g, and net holdings of foreign bonds, B_t^f. That is, the loans among domestic households still cancel out. Therefore, we can replace B_t and B_{t-1} by $B_t^g + B_t^f$ and $B_{t-1}^g + B_{t-1}^f$, respectively, to get

$$C_t + (B_t^g – B_{t-1}^g)/P + (B_t^f - B_{t-1}^f)/P + (M_t-M_{t-1})/P + K_t–K_{t-1} = \tag{17.2}$$

$$Y_t + r\cdot(B_{t-1}^g + B_{t-1}^f)/P – \delta K_{t-1} + (V_t–T_t)/P.$$

Next we need the government's budget constraint, which comes from an extension of equation (14.1) from chapter 14:

(17.3) $\quad G_t + V_t/P + r\cdot(B^g_{t-1}/P) + (H_t\text{-}H_{t\text{-}1})/P = T_t/P + (B^g_t - B^g_{t-1})/P + (M_t\text{–}M_{t\text{-}1})/P$.

The left-hand side has uses of funds, which include purchases of goods and services, G_t, transfers, V_t/P, and interest payments on government bonds, $r\cdot(B^g_{t-1}/P)$. The new item is the change in real international currency, $(H_t\text{-}H_{t\text{-}1})/P$. We are thinking of the central bank—a part of the government—as having to expend resources to acquire international currency. Thus, we are neglecting the fact that some governments, such as the United States and the European Union, can print dollars or euros that would be held as currency by central banks. The right-hand side of equation (17.3) has sources of funds, which are taxes, T_t/P, public debt issue, $(B^g_t - B^g_{t-1})/P$, and creation of domestic money, $(M_t\text{–}M_{t\text{-}1})/P$.

We can use equation (17.3) to substitute out

$$(B^g_t - B^g_{t-1})/P + (M_t\text{–}M_{t\text{-}1})/P = G_t + (V_t - T_t)/P + r\cdot(B^g_{t-1}/P) + (H_t\text{-}H_{t\text{-}1})/P$$

on the left-hand side of equation (17.2). If we cancel out the term $(V_t - T_t)/P + r\cdot(B^g_{t-1}/P)$, which appears on both sides, we get

$$C_t + (B^f_t - B^f_{t-1})/P + K_t\text{–}K_{t\text{-}1} + G_t + (H_t\text{-}H_{t\text{-}1})/P = Y_t - \delta K_{t\text{-}1} + r\cdot(B^f_{t-1}/P).$$

Finally, note that $K_t\text{–}K_{t\text{-}1}$ is net domestic investment, and $K_t\text{–}K_{t\text{-}1} + \delta K_{t\text{-}1}$ is gross domestic investment, I_t. Therefore, if we move $\delta K_{t\text{-}1}$ from the right-hand side to the left-hand side, we can replace $K_t\text{–}K_{t\text{-}1} + \delta K_{t\text{-}1}$ by I_t. The result is the budget constraint for the home country when it interacts with the rest of the world:

Key equation (budget constraint for open economy):

(17.5) $\quad C + I + G + (B^f_t - B^f_{t-1})/P + (H_t\text{-}H_{t\text{-}1})/P = Y + r\cdot(B^f_{t-1}/P).$

For convenience, we dropped the time subscripts where no confusion results. Notice that, for a closed economy, where $B^f = 0$ and $H = 0$, we have the familiar condition $C + I + G = Y$. An open economy has three new terms. On the left-hand side, we have the changes in net real foreign bonds, $(B_t^f - B_{t-1}^f)/P$,[2] and international currency, $(H_t-H_{t-1})/P$. On the right-hand side, we have the real interest income on the net quantity of foreign bonds, $r \cdot (B_{t-1}^f/P)$.

The left-hand side of equation (17.5) has uses of funds for the home country: consumption, C, gross domestic investment, I, government purchases of goods and services, G, and the accumulation of net real foreign bonds and international currency, $(B_t^f - B_{t-1}^f)/P + (H_t-H_{t-1})/P$. The term $(B_t^f - B_{t-1}^f)/P + (H_t-H_{t-1})/P$ is called **net foreign investment**, because it represents the net investment by the home country in the rest of the world. More generally, the assets include ownership of capital as well as bonds. Typically, additions to international currency are a minor part of net foreign investment.

The right-hand side of equation (17.5) has sources of funds for the home country: real GDP, Y, plus the real interest income received on the net holdings of foreign bonds, $r \cdot (B_{t-1}^f/P)$.[3] The last term is the real **net factor income from abroad**.[4] This term is the real income from ownership of assets (bonds or capital) in foreign countries net of the

[2] We assumed that all foreign bonds were held by households. However, equation (17.5) still applies if the home government holds foreign bonds or borrows from foreigners. The variable B^f represents the net position of the home country—government and households—with respect to holdings of foreign bonds.
[3] We are neglecting transfer payments, such as foreign aid, from one country to another.
[4] More generally, the net factor income from abroad also includes labor income—wage income of home country residents working abroad less wage income of foreign residents working in the home country. We have assumed that all home country residents work in the home country and all foreign residents work abroad. For the United States, the net labor income from abroad is small compared with GDP. For countries that export lots of workers to other places, such as El Salvador, Mexico, and Turkey, the net labor income from abroad is large in relation to GDP. Countries such as Germany and the Persian Gulf states, which import many "guest workers" from other countries, have a sizable negative net labor income from abroad.

real income going to foreigners from their ownership of assets (bonds or capital) in the home country.

The sum of real GDP and the real net factor from abroad—that is, the right-hand side of equation (17.5)—is the real gross national product (GNP). The GDP equals the total payments to factors of production (labor and capital) located in the home country. The GNP adds in the net payments to factors located abroad. In our analysis, these net factor payments consider only amounts paid to capital. More generally, the net factor income from abroad and, hence, GNP would also include the net payments to labor located abroad (see footnote 4).

We can use equation (17.5) to get the conditions for the **balance of international payments**. The right-hand side of the equation has the total real income of domestic residents, which is the real GNP, $Y + r\cdot(B^f_{t-1}/P)$. The left-hand side includes the total real expenditure on goods and services, $C+I+G$. If we subtract total real expenditure from both sides of the equation, we get

Key equation (balance of international payments):

$$(B^f_t - B^f_{t-1})/P + (H_t - H_{t-1})/P = Y + r\cdot(B^f_{t-1}/P) - (C+I+G). \quad (17.6)$$

The right-hand side of equation (17.6) is called the **current-account balance**. If the home country's real GNP, $Y + r\cdot(B^f_{t-1}/P)$, is greater than its real expenditure on goods and services, $C+I+G$, the home country is running a surplus with respect to the rest of the world, that is, the current account is in surplus. Conversely, if real GNP is less than real expenditure, the current account is in deficit. The left-hand side of equation (17.6) shows that the current-account balance equals the net increase in the home country's holdings of international assets—net foreign bonds and international currency. In other words, the current-account balance equals net foreign investment.

Exports are the goods and services produced domestically that are sold to foreign residents, and imports are the goods and services produced abroad that are bought by domestic residents. The difference between exports and imports, or net exports, is called the **trade balance**. The trade balance equals the goods and services produced domestically, which is the real GDP, Y, less the total of real domestic expenditure, $C+I+G$. That is, the goods and services produced at home but not purchased at home must be moving on net—as net exports—to foreign residents. The trade balance is positive or in surplus if exports exceed imports and negative or in deficit if imports exceed exports.

Equation (17.6) implies

current-account balance = trade balance + net factor income from abroad.

$$(B_t^f - B_{t-1}^f)/P + (H_t - H_{t-1})/P = Y - (C+I+G) + r\cdot(B_{t-1}^f/P) \tag{17.7}$$

Thus, the current-account balance, $(B_t^f - B_{t-1}^f)/P + (H_t - H_{t-1})/P$, differs from the trade balance, $Y - (C+I+G)$, by the net factor income from abroad, $r\cdot(B_{t-1}^f/P)$. If we had allowed for international transfers, such as foreign aid, we would subtract the net real transfers to foreigners from the right-hand side of equation (17.7). That is, the current-account balance is reduced by net real transfers from the home country (households or government) to the rest of the world.

Consider a country that has experienced a current-account surplus in the past and has therefore built up a net real asset position with respect to the rest of the world, B_t^f/P. The real net factor income from abroad, $r\cdot(B_{t-1}^f/P)$, is greater than zero, and equation (17.7) says that the current-account balance is greater than the trade balance. In other words, the country can have a zero current-account balance even if its imports exceed its exports, so that the trade balance is in deficit. The country pays for its excess imports with its net factor income from abroad.

We can get another perspective on the balance of international payments by thinking about the home country's saving and investment. We can rearrange the right-hand side of equation (17.7) to get

$$(B_t^f - B_{t-1}^f)/P + (H_t - H_{t-1})/P = Y + r\cdot(B_{t-1}^f/P) - (C+G) - I.$$

The first two terms on the right-hand side give real GNP, $Y + r\cdot(B_{t-1}^f/P)$. If we subtract depreciation, δK, we get real net national product (NNP), which equals real national income (see section II.B of chapter 2). Thus, if we subtract δK from $Y + r\cdot(B_{t-1}^f/P)$ and also subtract δK from I on the far right, we get

$$(17.8)\quad (B_t^f - B_{t-1}^f)/P + (H_t - H_{t-1})/P = Y + r\cdot(B_{t-1}^f/P) - \delta K - (C+G) - (I - \delta K).$$

The difference between real national income, $Y + r\cdot(B_{t-1}^f/P) - \delta K$, and the consumption by households and government, $C+G$, is the total real saving by households and government.[5] The right-hand side of equation (17.8) is the difference between this real national saving and net domestic investment, $I - \delta K$. Hence, the current-account balance can be expressed as[6]

Key equation (current-account balance, saving, and investment):

(17.9) real *current-account balance = real national saving – net domestic investment*.

[5] Recall that we have neglected public investment and have thought of all of the government's purchases of goods and services, G, as consumption.
[6] This result holds even when the government carries out public investment. In that case, domestic investment is the total of private and public investment in the home country.

The current account is in surplus when national saving is greater than net domestic investment, and the current account is in deficit when national saving is less than net domestic investment.

II. History of the U.S. Current-Account Balance

Figure 17.1 shows the ratio of the nominal U.S. current-account balance to nominal GDP (GNP before 1929) from 1820 to 2004. (The graph excludes the Civil War years, 1861-65.) The current account was typically in deficit before 1897, averaging -0.5% of GNP from 1820 to 1896. The current account was in surplus for most years from 1897 to 1976. The surplus averaged 0.7% of GNP from 1897 to 1914, then rose to 4.5% with the substantial lending to foreign allies during World War I. The surplus averaged 1.2% of GNP for the 1920s and 0.4% of GDP for the 1930s. During World War II, the current account was in surplus (1.3% of GDP) in 1940-41 before the United States entered the war. After the U.S. entrance, the current account shifted to deficit from 1942 to 1945, averaging -0.6% of GDP.

The current account moved to a large surplus in the immediate post-World War II years, averaging 2.8% of GDP for 1946-47. Then the surplus averaged 0.4% of GDP from 1948 to 1976. The current account was roughly balanced from 1977 to 1982, then shifted to deficit, reaching -3.2% of GDP in 1986-87. The current account returned to balance in 1991, then shifted again to deficit, reaching the remarkably high average magnitude of -4.4% of GDP from 2000 to 2004. In the prior U.S. history, the only ratio of comparable size was in 1836.[7]

[7] For the entire U.S. history, the largest deficit in relation to GNP was probably around -6% in 1816, following the end of the War of 1812. However, these early data are not very reliable.

The upper part of Figure 17.2 shows the ratios of nominal exports and imports to nominal GDP (GNP before 1929). These ratios illustrate the changing significance of international trade for the U.S. economy. The ratios averaged over 9% before the Civil War, 1820-60,[8] 6-7% for 1866-1914, and only 4-5% in the interwar period, 1921-40. After World War II, the ratios of exports and imports to GDP rose sharply. The export ratio increased from around 4% in 1950 to over 10% by the early 1990s, then remained relatively stable. The import ratio expanded from about 4% in 1950 to 15% by 2004.

The lower part of Figure 17.2 shows the ratio of the trade balance—nominal exports less nominal imports—to GDP or GNP. The variations in this ratio account for the main movements in the current-account ratio, shown in Figure 17.1. In the post-World War II period, the ratio of the trade balance to GDP fell from a large surplus immediately after World War II to near zero until the early 1980s. The trade deficit reached 3% of GDP in the mid 1980s, then returned to near zero at the start of the 1990s. Since then, the trade deficit has been large, rising to 4-5% of GDP in 2000-04. Thus, the large trade deficit accounts for the large current-account deficit in these years.

Figure 17.3 shows the ratio to GDP of the net factor income from abroad from 1929 to 2004. These data include the net labor income from abroad, but the main part is capital income. The net factor income from abroad climbed from 0.4% of GDP after World War II to a peak of 1.4% in 1980. Then the shift to a current-account deficit (Figure 17.1) meant that foreigners were accumulating net claims on the United States. This development led to a fall in the net factor income from abroad—to around 0.5% of GDP by the end of the 1980s.

[8] The less accurate earlier data indicate that the ratios were much higher before 1820. The ratios of exports and imports to GNP were 20-30% between 1795 and 1807.

As mentioned, in 2000-04, current-account deficits averaged 4-5% of GDP. The cumulative effect of these and earlier deficits meant that the estimated net international asset position of the United States by the end of 2003 was not only less than zero but staggeringly large in magnitude: $2.7 trillion or 26% of nominal U.S. GDP. Yet, surprisingly, the estimated net flow of capital income from abroad remained greater than zero. The explanation is that rates of return on U.S. capital held abroad were high, reflecting strong profits on direct investments and other assets in foreign countries. In contrast, rates of return on foreign capital held in the United States were much lower. One reason is that a significant part of the foreign holdings were in low-yield U.S. Treasury bonds. Most of these securities were held by foreign central banks, notably in Asia. Obviously, this situation is a good deal for the United States. The question is why are foreigners willing to hold so much wealth in low-yield U.S. securities—and will they continue to do this forever?

Figure 17.4 shows from 1929 to 2004 the last part of the current-account balance, the net U.S. transfers to foreigners. This item, not included in equation (17.7), enters negatively into the U.S. current-account balance. Net U.S. transfers abroad include official foreign aid and private transfers. The net U.S. transfers abroad were greater than zero in almost all years since 1929, averaging 0.5% of GDP. The positive spike in the late 1940s—greater than 2% of GDP—represents the large transfers to U.S. World War II allies. The negative spike in 1991—around 1% of GDP—reflects the payments to the United States from Saudi Arabia and other Persian Gulf countries allied with the United States during the Gulf War.

III. Real Business Cycles and the Current-Account Balance

In chapters 7 and 8, we used the real business cycle model to show how shocks to the technology level, *A*, generate economic fluctuations in a closed economy. We now extend these results to an open economy. To simplify the analysis, we neglect the extension from chapter 8 to allow for variability in the capital utilization rate, κ. In this case, we can represent capital services, κK, just by the capital stock, *K*. Therefore, the production function is

(17.10) $$Y = A \cdot F(K, L).$$

As usual, we treat *K* as fixed in the short run.

Start with a summary of our previous results on how an increase in the technology level, *A*, affects a closed economy. An increase in *A* raises the marginal products of labor, MPL, and capital, MPK. The rise in the MPL raises the demand for labor and leads, thereby, to increases in the real wage rate, *w/P*, and the quantity of labor, *L*. The rise in the MPK raises the demand for capital services and leads, accordingly, to a rise in the real rental price, *R/P*. By assumption, the quantity of capital services does not change in the short run. Therefore, the increase in *A* raises *Y* partly because of the direct effect of *A* in the production function in equation (17.10) and partly because of the increase in *L*. We also know that the expansion of *Y* shows up partly as more consumption, *C*, and partly as more gross domestic investment, *I*.

Consider how the real interest rate, *r*, is determined in a closed economy. With a fixed utilization rate for capital, κ, the rate of return on capital equals the real rental price, *R/P,* minus the depreciation rate, δ. Since the market for capital services clears, R/P equals the marginal product of capital, MPK. The real rate of return on capital has to equal the real interest rate on bonds, r. Therefore,

$$r = MPK - \delta. \quad (17.12)$$

Since the increase in the technology level, A, raises MPK, we find that r rises.

Suppose that the home country begins with the technology level A_1. We can use our familiar market-clearing conditions for a closed economy to determine the quantities—Y_1, L_1, C_1, I_1—and prices—$(w/P)_1, (R/P)_1, r_1$. One condition that has to be satisfied for a closed economy is that the real GDP equal the total of consumption, gross domestic investment, and government purchases of goods and services:

$$Y_1 = C_1 + I_1 + G_1. \quad (17.11)$$

We will be particularly interested in the real interest rate, r_1. Define the rate r_1 to be the real interest rate that would apply when the economy is closed and the technology level is A_1.

What happens if this closed economy suddenly gets access to the world credit market? We assumed earlier in this chapter that the world credit market operates effectively as a single market. Consequently, the real interest rates in the home country will have to be brought into equality with those in the rest of the world.

Suppose, to begin, that the real interest rate in the rest of the world, r^f, happens to equal r_1. In this case, the opening up to the world credit market would not change the real interest rate for the home country's households and government. Therefore, decisions about working, consuming, saving, investing, and so on would be the same as before. In particular, the country would have the same real GDP, Y_1, consumption, C_1, gross domestic investment, I_1, and so on. Therefore, the condition $Y_1 = C_1 + I_1 + G_1$

would continue to hold in equation (17.11). We see, accordingly, from equation (17.8) that the current-account balance would be zero.[9] Although the home country has the opportunity to borrow and lend on the world credit market, this option would not be exercised under the assumptions that we made.

To see the role of the world credit market, assume now that the technology level in the home country improves to the higher value A_2. We know for a closed economy that the market equilibrium dictates a higher real interest rate, r_2. This conclusion would also hold for the world economy if we had assumed that the improvement in technology applied to the whole world. That is, the world is a closed economy. Suppose, however, that the improvement in technology applies only at home, not to the rest of the world.

If the home economy is a negligible part of the world economy, we can treat the foreign real interest rate, r^f, as independent of outcomes at home.[10] In this case, if the home country's real interest rate rose to r_2, the home real interest rate would be higher than the foreign rate, r^f, which does not change. But then foreigners would want to do all of their lending in the home country, whereas domestic residents would want to do all of their borrowing abroad. To see how markets come into equilibrium, we have to make our model a little more complicated than before.

Equation (17.12) says that the home country's real interest rate, r, must equal MPK – δ. The rise in the home country's technology level, A, raises the MPK and, therefore, tends to raise r above the foreign rate, r^f. To get an equilibrium where r^f does

[9] We are assuming that the home country starts with a zero net asset position with respect to the rest of the world, $B^f = 0$. We are also neglecting acquisitions of international currency, H—usually these currency flows are a minor part of the story.

[10] If the home country is the United States, its economy would be large enough to have a noticeable impact on worldwide real interest rates. We could modify our analysis accordingly, but the same basic ideas still apply.

not change and r equals r^f, the rate of return on capital at home, MPK – δ, has to come back down to its initial value. The problem is that, if we treat the stock of capital at home, K, as fixed in the short run—that is, if we neglect the contribution of the current flow of net investment to the stock of capital—there is no way for the MPK to come down.[11]

One way that economists resolve this problem is to bring in **adjustment costs for investment**. If these costs rise with the amount of investment, then a large enough quantity of gross domestic investment, I—financed by loans from foreigners—would bring the rate of return on domestic investment down to equal the foreign real interest rate, r^f.

To study the main effects on the current-account balance, we do not have to resolve these issues in detail. The main point is that the increase in the home marginal product of capital, MPK, induces a large volume of gross domestic investment, I. Hence, from the perspective of equations (17.8) and (17.9), I rises compared to national saving. The rise in the difference between I and national saving equals the increase in the current-account deficit. However, if adjustment costs for investment are large, it is possible that I would rise by less than national saving. In that case, the current account would move toward surplus, rather than deficit.

We see from equation (17.7) that the movements in the current-account balance reflect changes in the trade balance. Consider the case in which an increase in the technology level, A, moves the current-account balance toward deficit. The mechanism for this change is an increase in imports of goods and services by more than the increase

[11] The MPK would be affected by variations in labor input, L, and the utilization rate for capital, κ. However, since L and κ tend to rise in this example, the MPK tends to increase further. Hence, these adjustments would not bring r back down to r^f.

in exports. That is, net imports rise. The increase in net imports provides the goods and services that allow gross domestic investment to rise by more than national saving.

For the case where adjustment costs for investment are small, the prediction is that the current-account balance and the trade balance would be counter-cyclical. That is, the current account and the trade balance would move toward deficit during an economic boom and toward surplus during a recession.

To check out these predictions, Figures 17.5 and 17.6 examine the cyclical behavior of U.S. real exports and real imports. As usual, we look at cyclical parts by filtering out trends. We see from the graphs that the cyclical parts of the logs of real exports and real imports have similar variability and are each more variable than the cyclical part of real GDP. From 1954.1 to 2004.3, the standard deviations were 0.048 for the log of real exports and 0.047 for the log of real imports, compared to 0.016 for the log of real GDP. Exports and imports are both procyclical—the correlations of the cyclical parts with the cyclical part of real GDP are 0.38 for exports and 0.71 for imports. Thus, imports are substantially more procyclical than exports. This finding suggests that the trade balance would be countercyclical, because imports would rise more than exports in a boom and fall more than exports in a recession. We examine the trade balance directly in the next figure.

Figure 17.7 shows the cyclical pattern of the ratio of the nominal trade balance to nominal GDP. The conclusion, as anticipated, is that the trade-balance ratio is somewhat countercyclical. The correlation of the cyclical part of the ratio with the cyclical part of real GDP is -0.34.

It may be surprising that, at least for the United States, the current-account balance typically gets worse—moves toward deficit—in good times and improves in bad times. To understand this result, think of a recession as a temporary shortfall of real GDP, Y, from normal (or trend). We know from chapter 6 that households seek to maintain their consumption when real income falls temporarily. Hence, household real saving tends to fall in a recession. If foreign countries are not suffering from a recession at the same time, we would have expected households in the home country to borrow from foreigners to maintain consumption. This perspective suggests that the current account would move toward deficit during a recession. In accordance with this view, equation (17.9) says that a decline in real national saving moves the current-account balance toward deficit.

The reason we got the opposite prediction for the current-account balance is that, in a recession, the decrease in the marginal product of capital, MPK, reduces net domestic investment, $I - \delta K$, by even more than the fall in real national saving. Consequently, we see from equation (17.9) that the current account moves toward surplus.

We get different results if the underlying shocks affect household real income but not the marginal product of capital, MPK. One possibility is a harvest failure—if this failure decreases the agricultural goods currently available but does not change the returns to investment in agriculture (that is, does not affect the MPK). Another case that has similar effects on the current-account balance is a temporary increase in the government's purchases of goods and services, G, such as in wartime. However, unlike a harvest failure, a temporary rise in G is usually not associated with a recession.

To assess these cases formally, suppose that, in a recession, real GDP, Y, falls. Since the change in Y is temporary, the income effect on consumption, C, is weak. Hence, C tends to fall by less than Y, and real national saving declines. (There are no intertemporal-substitution effects, because the home real interest rate, r, is fixed—it has to equal the foreign rate, r^f, which has not changed.) We also assume that the marginal product of capital, MPK, has not changed. Therefore, net domestic investment, $I - \delta K$, does not change. Equation (17.9) gives the effect on the current-account balance. This balance moves toward deficit, because real national saving falls, and domestic net investment does not change.

We also see from equation (17.8) that the same effect arises if government purchases, G, rise temporarily. Since the income effect is small, households tend to reduce consumption, C, by less than the rise in G. Therefore, the current-account balance moves toward deficit. We can think of the government as financing its temporarily high purchases by borrowing from foreigners. An example is the borrowing of prospective U.S. allies from the United States at the starts of World Wars I and II, before the United States entered the wars.

Examples of international borrowing and lending

One empirical study that supports our analysis involves wheat harvests in Australia from 1931 to 1985 (see John Scoggins [1990]). Poor harvests reduced real GDP in Australia but had little impact on the rest of the world. Effects on investment in Australia were minor. Therefore, as the model predicts, poor harvests led to increases in Australia's trade deficit. That is, harvest failures motivated Australians to borrow from foreigners to maintain levels of consumption.

A similar analysis applies to the massive harvest failure in Poland in 1978-81. In response to the temporary drop in real income, Poland's foreign debt in 1981 reached $25 billion, roughly half of the country's GDP.[12] Thus, as in Australia, a harvest failure motivated Poles to borrow from foreigners to maintain consumption.

For effects of government purchases, we can consider the linkage between war and the U.S. current-account balance (see Figure 17.1). From 1915 to 1917, before the United States entered World War I, the prospective allied countries borrowed large sums from the United States to finance their temporarily high government purchases. However, once the United States entered the war in 1917, the U.S. current-account surplus diminished—because the United States had to pay for its own government purchases. Similarly, the United States ran a substantial current-account surplus in 1940-41, before it entered World War II. After the United States entered the war and U.S. government purchases soared, the current account shifted to a small deficit. (One difficulty with a nearly global war is that there are few non-combatants from whom to borrow. We have to

[12] The data on foreign debt for Poland (and for Mexico and Brazil, mentioned later) are from Morgan Guaranty Trust, *World Financial Markets*, February 1983, and Organization of American States, *Statistical Bulletin of the OAS*, January-June 1982.

think then about the effects of temporarily high government purchases in a closed economy.)

Mexico's discovery of oil in the early 1970s brings in other forces that affect international borrowing. By 1974, Mexico's oil prospects were great, but a significant amount of production had yet to materialize. Investment was high in oil and related industries. Consumption—by households and government—rose because of the increase in prospective real income. Since real GDP was not yet high, Mexico was motivated to borrow from foreigners to pay for current investment and consumption. Consequently, Mexico's foreign debt increased from $3.5 billion or 9% of GDP in 1971 to $61 billion or 26% of GDP in 1981.

We found before that shifts in investment were the key to understanding the cyclical behavior of the U.S. current-account balance. The current account tended to be in deficit in good times because investment was high at those times. This perspective also explains why developing countries with strong opportunities for growth and investment tend to run current-account deficits. For example, from 1971 to 1980, Brazil's real per capita GDP rose by 5% per year. During this period, Brazil's foreign debt rose from $6 billion, or 11% of GDP, to $55 billion, or 22% of GDP.

For an earlier example of a developing country that borrowed heavily from abroad, consider the United States. In 1890, the foreign debt reached $2.9

billion, or 21% of GNP. Recall from Figure 17.1 that the United States ran a current-account deficit for most years prior to 1897 (a pattern that applies back to at least 1830).

IV. The Current-Account Deficit and the Budget Deficit

In chapter 14, we studied how budget deficits affect the economy. Think about a cut in year *t*'s real taxes, T_t/P, corresponding to an increase in the real public debt, $(B_t^g - B_{t-1}^g)/P$. We assume that taxes are lump sum, although we could also consider effects from distorting taxes, as we did in chapter 13. We also assume that the time path of government purchases, *G*, does not change.

In the Ricardian approach, discussed in chapter 14, we found cases in which a deficit-financed cut in year *t*'s real taxes, T_t/P, did not affect the present value of taxes paid by households. Under these circumstances, households did not change consumption, C_t. Since the cut in T_t/P raised real disposable income and C_t did not change, real private saving in year *t* went up by the full amount of the tax cut. The increase in real private saving therefore completely offset the reduction in real public saving, corresponding to the budget deficit. Hence, the total of real private and public saving—or real national saving—did not change.

In a closed economy, we found in the Ricardian case that a budget deficit had no effect on macroeconomic variables, including real GDP, *Y*, consumption, *C*, gross investment, *I*, and the real interest rate, *r*. The reason was that households saved the entire cut in year *t*'s real taxes, T_t/P. If we allow for an open economy, the important point is that the budget deficit does not affect real national saving. Therefore, we see from equation (17.9) that the current-account balance would not change. Since households in the home country save the full amount of the tax cut, they willingly absorb all of the additional government bonds issued by the government. Consequently, the home country does not borrow more from the rest of the world, and the current-account balance does not change. Thus, in the Ricardian case, a budget deficit does not create a current-account deficit.

The conclusions are different if households do not save the full amount of the tax cut, that is, if a budget deficit changes real national saving, S. For example, we mentioned in chapter 14 that finite-lived households might feel wealthier when the government cuts taxes and runs a budget deficit. In this case, consumption, C, rises. Therefore, real private saving increases by less than the tax cut, and real national saving, S, declines. Equation (17.9) shows that the current-account balance moves toward deficit. That is, the home country borrows more from foreigners to pay for the rise in consumption at home. Hence, in this case, a budget deficit leads to a current-account deficit.

When budget and current-account deficits occur at the same time, an economy is said to be suffering from **twin deficits**. Economists applied this label to the U.S. situation of the mid 1980s, when budget deficits were large and the current-account deficit gradually widened (see Figure 17.1). However, the current-account deficit disappeared at the beginning of the 1990s, despite the continued presence of budget deficits. Moreover, the current-account deficit rose again late in the 1990s even though the budget shifted toward surplus. However, twin deficits did appear again in the United

States in 2002-04. Thus, although twin deficits sometimes occur, they are not a regular feature of the U.S. economy or, it turns out, of other economies.

An empirical study by Paul Evans (1988) considered whether exogenous movements in budget deficits cause current-account deficits. His sample was the post-World War II period for the United States, Canada, France, Germany, and the United Kingdom. The findings, consistent with the Ricardian view, were that budget deficits did not have a statistically significant causal influence on current-account deficits.

We should note that, even if budget deficits do not cause current-account deficits, twin deficits can arise as responses to some disturbances. Consider, for example, the effects from a temporary expansion of government purchases, G, as in wartime. We found before that a temporary rise in G tends to move the current account toward deficit. We also know from chapter 14 that an increase in G motivates the government to run a budget deficit in order to avoid a large temporary increase in taxes. Therefore, we predict that the current-account and budget deficits would move together in this case. However, we would not say that the budget deficit caused the current-account deficit. Rather, the two deficits move endogenously in the same direction in response to a common shock—it this case, the rise in government purchases.

V. The Terms of Trade

We have assumed, thus far, that the home country and the rest of the world produce goods that are physically identical. This assumption is unrealistic because countries—particularly small ones—tend to specialize in the production of different goods. For example, Chile produces a lot of copper, Brazil produces a lot of coffee, and

Saudi Arabia produces a lot of oil. Given these patterns of specialization, countries are affected substantially when the prices of their principal products change compared to prices of other goods. A particularly important price for the world economy is the oil price. All countries use oil and related petroleum products, but the production of oil is concentrated in a relatively small number of places. A rise in the price of oil, compared to other prices, is good for the relatively small number of oil producers but bad for other countries.

To keep things simple, pretend that the home country produces a single good that sells everywhere at the dollar price P, whereas the rest of the world produces another good that sells everywhere at the price $\overline{P}$. When the home country exports goods, it receives $\$P$ for each unit of goods exported. When the home country imports goods, it pays $\$\overline{P}$ for each unit of goods imported.

Consider the ratio $P/\overline{P}$. This ratio is called the **terms of trade**. The terms of trade has the units:

($ per home good)/($ per foreign good) = foreign good per home good.

That is, the terms of trade gives the number of units of foreign goods that can be imported for each unit of home good exported. If the terms of trade, $P/\overline{P}$, rises—or improves—the home country is better off because it gets more foreign goods in exchange for each unit of home goods. If $P/\overline{P}$ falls—or worsens—the home country is worse off because it gets fewer foreign goods in exchange for each unit of home goods.

A. The terms of trade and the current-account balance

In the real business cycle model, economic fluctuations were generated by shocks to the technology level, *A*. For a single country, changes in the terms of trade have effects that are similar to changes in *A*. An improvement in the terms of trade resembles an increase in *A*, whereas a worsening of the terms of trade resembles a decrease in *A*. To see how this works, we have to incorporate the terms of trade into our equations for the balance of international payments.

Equation (17.6) gave the current-account balance in real terms when all goods sold at the single dollar price, *P*. If we multiply that equation through by *P*, we get that the current-account balance in nominal terms is

nominal current-account balance = nominal GNP – nominal domestic expenditure.

$$(B_t^f - B_{t-1}^f) + (H_t - H_{t-1}) = PY + r\cdot B_{t-1}^f - P\cdot(C+I+G)$$

Note that nominal domestic expenditure is the amount spent on consumption, *PC*, gross domestic investment, *PI*, and government purchases, *PG*.

We have to modify this equation to bring in different prices for home and foreign goods. The simplest setting, which brings out the main results, is one where the home country exports all of its production, so that none of domestic output (real GDP) goes to the home country's consumption, investment, and government purchases. Instead, the home country imports foreign goods to provide for consumption, investment, and government purchases.

Given these assumptions, the nominal income from the real GDP of *Y* is *PY*, where *P* is the dollar price of home goods. The nominal expenditure on consumption, investment, and government purchases is $\overline{P}\cdot(C+I+G)$, where $\overline{P}$ is the dollar price of foreign goods. Therefore, the nominal current-account balance becomes

$$(17.12) \qquad (B_t^f - B_{t-1}^f) + (H_t - H_{t-1}) = PY + r\cdot B_{t-1}^f - \overline{P}\cdot(C+I+G).$$

nominal current-account balance = nominal GNP – nominal domestic expenditure.

If we divide equation (17.12) through by the price of foreign goods, $\overline{P}$, we get that the current-account balance in real terms is

real current-account balance = real GNP – real domestic expenditure.

$$(17.13) \qquad (B_t^f - B_{t-1}^f)/\overline{P} + (H_t - H_{t-1})/\overline{P} = (P/\overline{P})\cdot Y + r\cdot(B_{t-1}^f/\overline{P}) - (C+I+G).$$

Each term in equation (17.13) is a real value in the sense of being measured in units of foreign goods. The important difference from equation (17.6) is that Y (the real GDP measured in units of home goods) is multiplied by the terms of trade, $P/\overline{P}$. If $P/\overline{P}$ rises and Y does not change, the home country has more real income when measured in units of the foreign goods that can be purchased.

As before, we can express the results for the current-account balance in terms of real national saving and net domestic investment. In equation (17.13), if we subtract δK from real GNP, $(P/\overline{P})\cdot Y + r\cdot(B_{t-1}^f/\overline{P})$, and also from gross domestic investment, I, we can rearrange terms to get

$$(17.14) \quad (B_t^f - B_{t-1}^f)/\overline{P} + (H_t - H_{t-1})/\overline{P} = (P/\overline{P})\cdot Y + r\cdot(B_{t-1}^f/\overline{P}) - \delta K - (C+G) - (I - \delta K).$$

The term $(P/\overline{P})\cdot Y + r\cdot(B_{t-1}^f/\overline{P}) - \delta K$ is real national income (in units of foreign goods). The term $C+G$ is total domestic consumption by households and government. The difference between real national income and total domestic consumption is real national saving. Therefore, we have, as in equation (17.9), (17.15) real *current-account balance = real national saving – net domestic investment*.

Now we consider how an improvement in the terms of trade, $P/\overline{P}$, affects economic choices in the home economy. We still assume that the production function is given by

$$Y = A \cdot F(K, L). \tag{17.10}$$

We continue to assume that the capital stock, K, is fixed in the short run and that the capital utilization rate, κ, is constant. We also assume that the technology level, A, is fixed. Suppose for the moment that the quantity of labor, L, does not change. In this case, real GDP, Y, will not change in equation (17.10). Recall, however, that Y is real GDP in units of domestic goods.

Equation (17.14) shows that an increase in the terms of trade, $P/\overline{P}$, raises real national income, $(P/\overline{P})\cdot Y + r\cdot(B^f_{t-1}/\overline{P}) - \delta K$, for a given real GDP, Y. Households respond to the higher real income by increasing consumption, C. The response of C is greater the more permanent the rise in $P/\overline{P}$. In any event, the rise in C is typically less than the increase in real national income, so that real national saving rises. Thus, on this count, the current-account balance moves toward surplus in equation (17.15).

Now we have to consider the effect of a rise in the terms of trade, $P/\overline{P}$, on net domestic investment, $I - \delta K$. To assess this response, we have to consider how the rate of return on capital in the home country depends on $P/\overline{P}$. As before, an increase in the capital stock, K, by one unit raises output, Y, by MPK units. However, the price of K is now $\overline{P}$, the foreign price level, because investment comes from imports of foreign goods. The price of Y is still P, that is, domestic output sells at the domestic price level. Thus, a dollar outlay of $\overline{P}$ for one unit of capital yields a dollar flow of returns of

P·MPK. The gross rate of return on capital is the ratio of P·MPK to $\overline{P}$, that is, $(P/\overline{P})$·MPK. The net rate of return subtracts the depreciation rate, δ:

$$\textit{net rate of return on capital} = (P/\overline{P})\cdot MPK - \delta.$$

The difference from before is that the MPK is multiplied by the terms of trade, $P/\overline{P}$.

For a given MPK, an improvement in the terms of trade, $P/\overline{P}$, raises the net rate of return on capital in the home country. Think about Saudi Arabia as the home country, and consider the effect from an increase in the relative price of oil. The Saudis would see an increase in the prospective rate of return on domestic investments in the oil business.

The effect of $P/\overline{P}$ on the rate of return on capital is analogous to the effect that we considered before from an increase in the technology level, A. In that case, the increase in the MPK raised the rate of return on capital. In our previous discussion, we noted that the response of investment would be large, greater than the increase in real national saving. The same conclusion applies here. Therefore, the current-account balance in equation (17.15) would move overall toward deficit.

Also as before, we can get different results for the current-account balance if we introduce adjustment costs for investment. If these costs are large, it is possible that an increase in the terms of trade, $P/\overline{P}$, would raise net domestic investment by less than real national saving. For example, if an increase in the relative price of oil is perceived as temporary, the response of investment would be small. Saudi Arabia would not find it worthwhile to invest a lot in additional drilling capacity in response to a rise in the oil price that is perceived to be short-lived. In this case, a rise in $P/\overline{P}$ would move the current-account balance toward surplus.

B. Variations in labor and real GDP

Up to now, we assumed that labor, L, did not vary in response to a change in the terms of trade, $P/\overline{P}$. Since capital input was also fixed in the short run, the real GDP, Y, did not change. To see how a rise in $P/\overline{P}$ affects L, we have to study the labor market.

Figure 17.8 shows labor demand and supply curves. Demanders of labor in the home country still pay the nominal wage rate, w, for labor and receive the nominal price, P, for their output. Therefore, the quantity of labor demanded, L^d, depends, as before, on the real wage rate, w/P. In other words, a rise in $P/\overline{P}$ does not shift the L^d curve—shown in red in the figure.

Suppliers of labor in the home country care about the nominal wage rate, w, measured relative to the price of consumer goods. Since we assumed that all of home consumption took the form of imports of foreign goods, the price of consumption goods is the foreign price, $\overline{P}$. Therefore, the real wage rate from the standpoint of workers is $w/\overline{P}$, not w/P. Notice that

$$w/\overline{P} = (w/P)\cdot(P/\overline{P}).$$

For a given w/P, an increase in the terms of trade, $P/\overline{P}$, means that $w/\overline{P}$ has gone up. An improvement in the terms of trade allows workers to buy more consumption goods for each hour worked.

A rise in the real wage rate for workers, $w/\overline{P}$, has, as usual, substitution and income effects on the quantity of labor supplied, L^s. The substitution effect between work and consumption motivates an increase in, L^s. The rise in $w/\overline{P}$ also means that workers are better off; hence, the income effect motivates a decrease in L^s. In our previous analysis of this type of situation in chapter 8, we assumed that the substitution effect dominated the income effect. As before, the dominance of the substitution effect is more likely when the change in $w/\overline{P}$ is temporary, that is, when the improvement in the terms of trade, $P/\overline{P}$, is temporary. In this case, the income effect is weak, and the substitution effect is likely to outweigh the income effect. In any event, our assumption in Figure 17.8 is that the substitution effect dominates the income effect. Therefore, an improvement in the terms of trade, $P/\overline{P}$, raises the quantity of labor supplied, L^s, at a given w/P. We show this effect as the rightward shift from the green curve, labeled L^s, to the blue one, labeled $(L^s)'$.[13]

We see from Figure 17.8 that the improvement in the terms of trade, $P/\overline{P}$, leads to an increase in the home country's labor input, L. We also see that the real wage rate measured in terms of domestically produced goods—that is, w/P—declines. This fall is necessary to induce the home country's producers to employ more labor. However, the real wage rate as viewed by workers, $w/\overline{P}$, increases. This rise is necessary to induce the home country's workers to work more.

The increase in labor input, L, raises real GDP, Y, through the production function in equation (17.10). Therefore, in equation (17.14), real national income, $(P/\overline{P})\cdot Y + r\cdot(B^f_{t-1}/\overline{P}) - \delta K$, rises partly because of the increase in the terms of trade, $P/\overline{P}$, and partly because of the induced increase in Y. The change in Y does not alter our main conclusions about the current-account balance. We still get that a rise in $P/\overline{P}$ tends to move the current account toward deficit because it raises net domestic investment, $I - \delta K$, by more than real national saving. Also as before, the inclusion of adjustment costs for

[13] More realistically, households also consume home-produced goods. In this case, the relevant price level for workers is a weighted average of $\overline{P}$ and P, and the relevant real wage rate is a weighted average of $w/\overline{P}$ and w/P. We still have that, for a given w/P, an increase in $P/\overline{P}$ raises the real wage rate as viewed by workers. Therefore, the analysis in Figure 17.8 continues to apply.

investment can reverse the results. Overall, our findings are that an improvement in the terms of trade, $P/\overline{P}$, has effects on real GDP and the current-account balance that are similar to those from an increase in the technology level, A.

Why was the U.S. current-account deficit so large in 2000-04?

We see from Figure 17.1 that the U.S. current-account deficit of 4-5% of GDP in 2000-04 is an outlier compared with the entire history since 1869. Why did the current-account deficit become so large? We cannot be sure of the answer, but we can offer suggestions. For further discussion, see Maurice Obstfeld and Kenneth Rogoff (2004).

We see from Figure 17.1 that the ratio of the U.S. current-account deficit to GDP built up steadily in the 1990s, while the U.S. economy was strong. The ratio of nominal investment to nominal GDP rose sharply, from 13% in 1991 to a peak of 18% in 2000. The increase in the ratio of the current-account deficit to GDP from near zero in 1991 to 4% in 2000 reflected the usual pattern whereby investment expanded by more than real national saving during an economic boom. This effect was probably more potent than usual because the growth of high-tech sectors—notably telecommunications and the Internet—had an especially strong effect on investment.

The economy went into a recession in 2001-02, partly because of the end of the technology boom in mid 2000 and partly because of the adverse consequences from the terrorist attacks of September 11, 2001. The ratio of nominal investment to GDP declined to 15% in 2002, and this change would, by itself, have reduced the ratio of the current-account deficit to GDP. However, the U.S. government had a large expansion in its purchases, partly for defense and other aspects of national security and partly for other programs. A substantial part of the increased federal purchases actually preceded the September 11th event, but the increase became more pronounced after September 11 (see Figures 12.1 and 12.2 in chapter 12). This expansion of government purchases likely explains why the current-account ratio stayed relatively steady in 2001-02 and then increased further to 5% in 2003-04. The growth in government expenditure, combined with the 2001-02 recession, also likely explains the return to budget deficits in 2002-04 (see Figure 14.3 in chapter 14.)

VI. Summing Up

Up to now, we studied a single closed economy. This neglect of the world economy is especially problematic with the opening up of international markets—that is, globalization—over the last fifty years. This chapter extends the model to allow for international trade in goods and services and for international borrowing and lending.

The current-account balance for the home country determines the change over time in net foreign assets. The current account is in surplus if real gross national product

(GNP) is greater than total spending on consumption, gross private domestic investment, and government purchases. Otherwise, the current account is in deficit. The GNP adds to GDP the real income on net foreign assets (the net factor income from abroad). We can express the current-account balance as the trade balance (exports less imports) plus the net factor income from abroad. We can also express it as the difference between real national saving and net domestic investment.

In the real business cycle model, a favorable technology shock typically raises net investment by more than real national saving. Hence, the current account balance moves toward deficit. Thus, the current-account balance is countercyclical. Equivalently, imports are more procyclical than exports, and the trade balance is also countercyclical. The predictions are different if a disturbance affects real national income but not the marginal product of capital. In this case, a favorable shock raises real national saving but leaves net investment unchanged. Thus, the current account moves toward deficit in response to a harvest failure or a temporary increase in government purchases, as in wartime.

In the Ricardian case, a budget deficit does not affect the current-account balance. However, budget and current-account deficits can occur together in response to certain shocks, such as movements in government purchases. Thus, twin deficits can arise even if budget deficits have no causal influence on current-account deficits.

The terms of trade is the ratio of export prices to import prices. Changes in the terms of trades have effects similar to those from technology shocks. An improvement in the terms of trade creates a current-account deficit if the increase in net domestic investment is larger than the rise in real national saving. Otherwise, an improvement in the terms of trade leads to a current-account surplus.

Questions and Problems

Mainly for review

17.1. Why is it infeasible for all countries to run a current-account deficit at the same time?

17.2. If a country runs a budget deficit must it also run a current-account deficit? How does the linkage between the two deficits depend on the relation between budget deficits and national saving?

Problems for discussion

17.x. A technology shock for a single country

Consider a temporary rise in the technology level, A, in the home country. Assume that no change occurs in the terms of trade.

a. What happens to the home country's labor, L, capital utilization, κ, and real GDP, Y?

b. What happens to the home country's consumption, C, and current-account balance?

17.x. Taxes and the current-account balance

Discuss the effects on a country's current-account balance from the following changes in tax rates:

a. A permanent increase in the tax rate on labor income, τ_w.

b. A temporary increase in the tax rate on labor income, τ_w.

c. A permanent increase in the tax rate on asset income, τ_w.

d. A temporary increase in the tax rate on consumption.

17.x. A change in the terms of trade

Suppose that the home country is Brazil, which produces a lot of coffee for export. We considered in the text a change in the terms of trade that stemmed from a disturbance in the rest of the world. Suppose, instead, that a temporary shock at home—say a failure of Brazil's coffee crop—raises the relative price of coffee. In this case, how does the disturbance affect Brazil's current-account balance?

17.x. Tariffs

Suppose that a small country levies a tariff on imports of a tradable good from abroad. If the good sells at price $\overline{P}$ abroad and the tariff rate is 10%, domestic residents pay the price $1.1 \cdot \overline{P}$ for each unit of the good. If the tariff is permanent, what are the effects on the home country's

a. consumption of tradable and nontradable goods?

b. production of tradable and nontradable goods?

c. current-account balance?

(Hint: do not forget the revenue that the government gets from the tariff.)

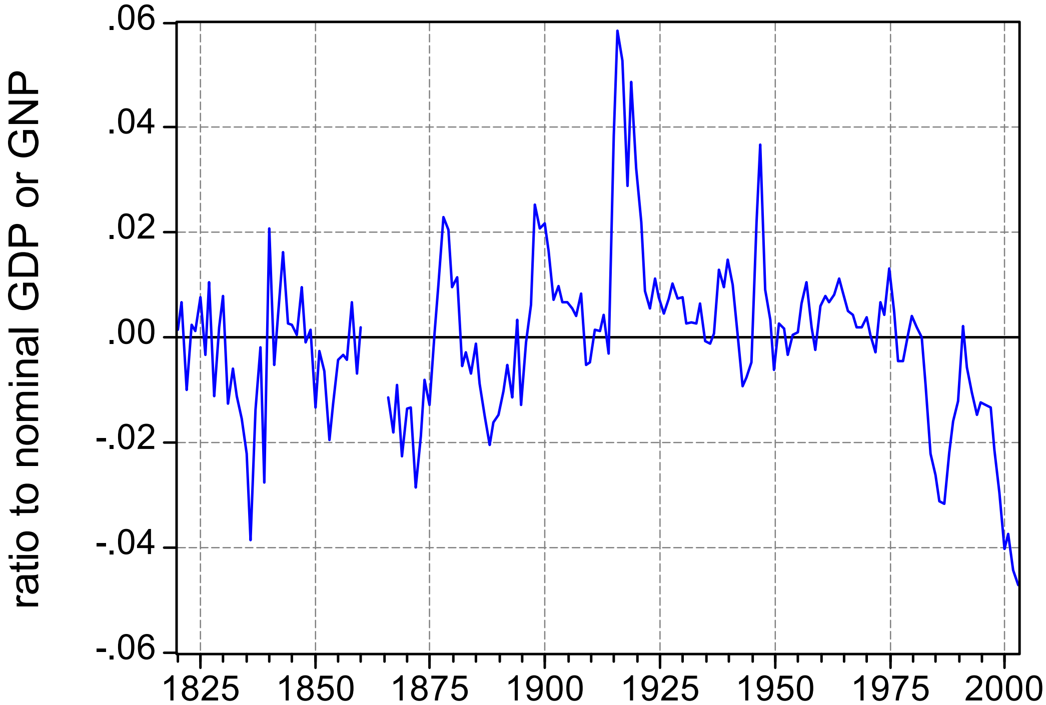


Figure 17.1

Ratio of U.S. Current-Account Balance to GDP, 1820-2004

The graph shows the ratio of the nominal current-account balance to nominal GDP (GNP before 1929). The data on the current-account balance are from U.S. Department of Commerce (1975) and the Bureau of Economic Analysis. (Numbers are omitted for the Civil War years, 1861-65.) The data on GDP and GNP are discussed in Figure 1.1 of chapter 1.

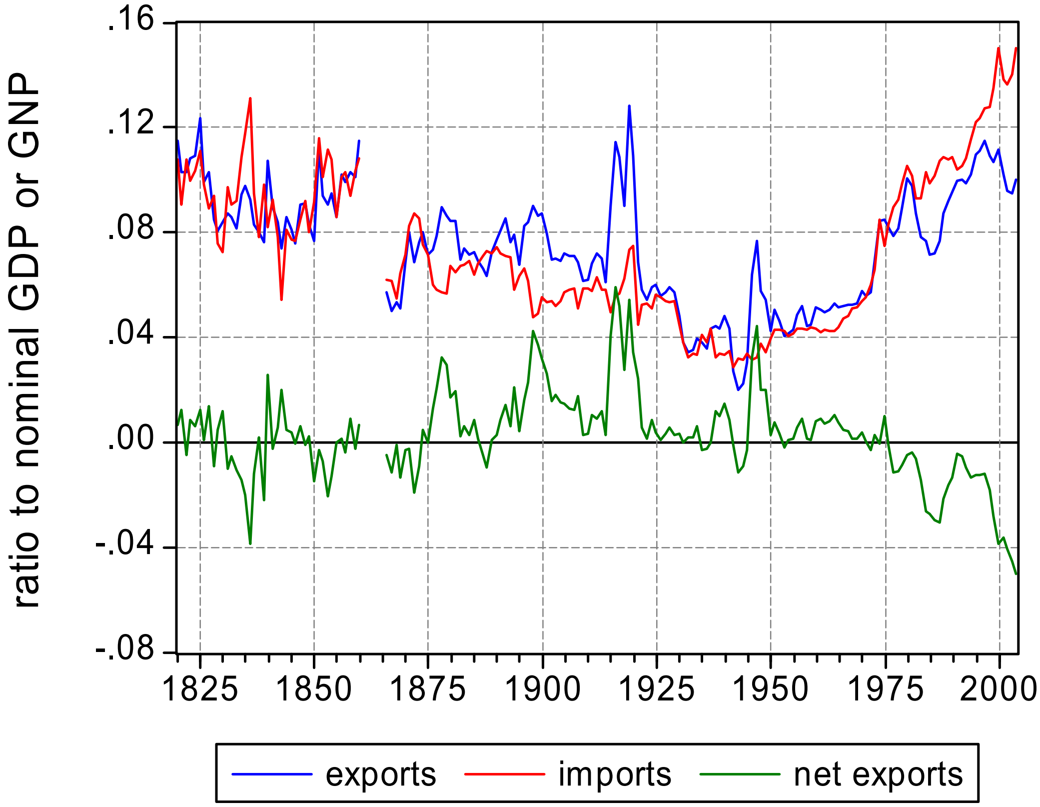


Figure 17.2

Ratios of Exports and Imports to GDP, 1820-2004

The upper curves show the ratios of nominal exports and imports of goods and services to nominal GDP (GNP before 1929). The lower curve shows the ratio of the trade balance (nominal exports less nominal imports) to nominal GDP or GNP. The data on exports and imports are from U.S. Department of Commerce (1975), Kendrick (1961), and the Bureau of Economic Analysis. (Numbers are omitted for the Civil War years, 1861-65.) The data on GDP and GNP are discussed in Figure 1.1 of chapter 1.

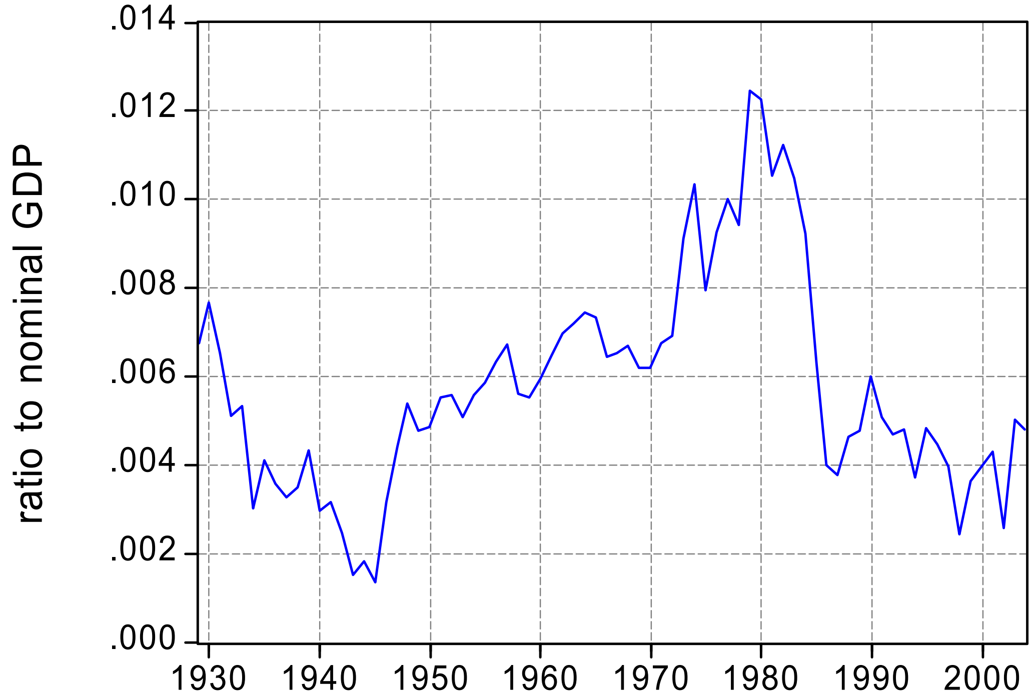


Figure 17.3

Ratio of Net Factor Income from Abroad to GDP, 1929-2004

Net factor income payments are income paid by foreigners to domestic capital and labor less income paid by domestic residents to foreign capital and labor. The graph shows the ratio of these net payments to nominal GDP. The data are from Bureau of Economic Analysis.

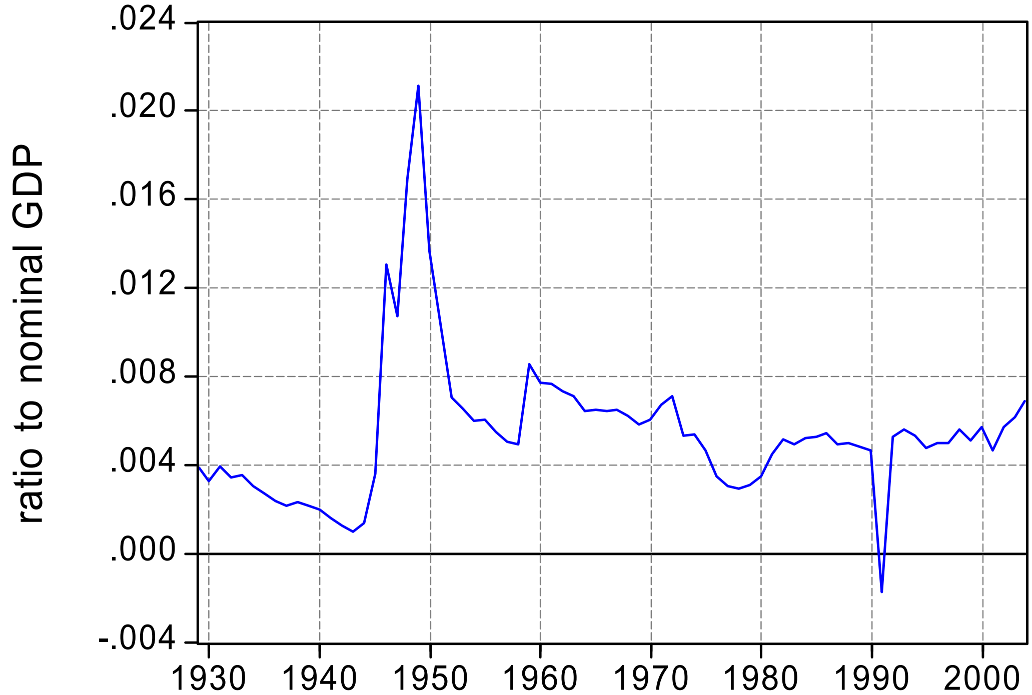


Figure 17.4

Ratio of Net U.S. Transfers Abroad to GDP, 1929-2004

The graph shows the ratio of net nominal transfers abroad to nominal GDP. The positive spike in the late 1940s represents the large transfers to U.S. World War II allies. The downward spike in 1991 reflects transfers to the United States from Saudi Arabia and other Gulf War allies. The data are from Bureau of Economic Analysis.

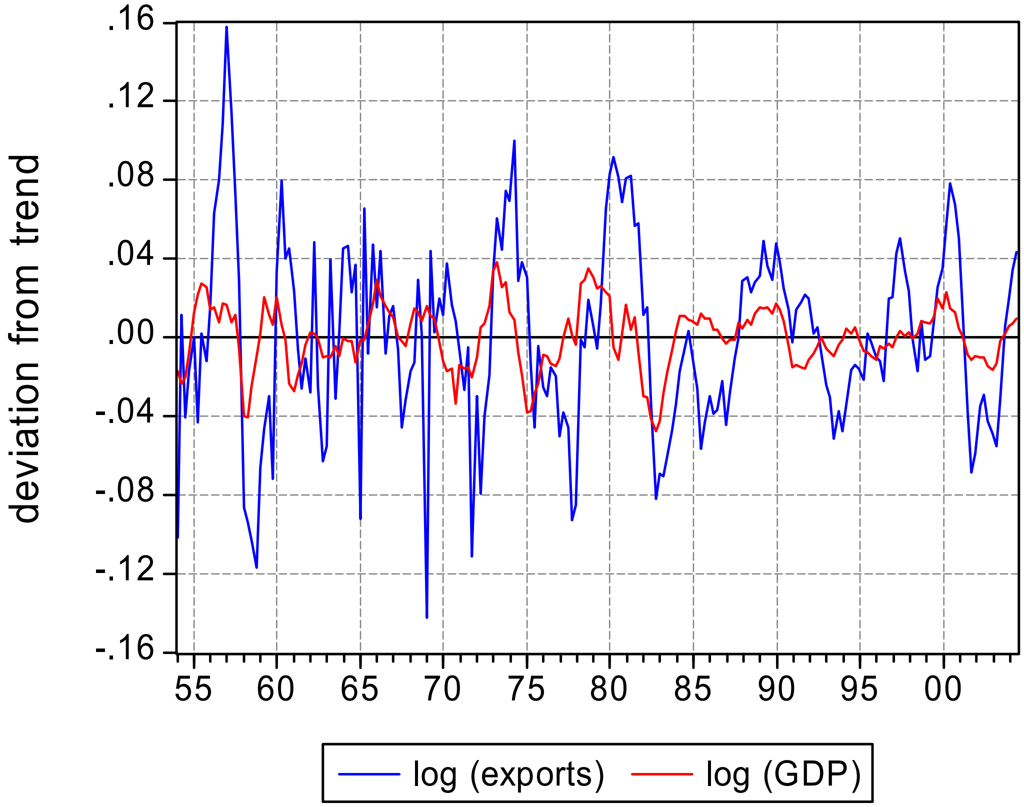


Figure 17.5

Cyclical Behavior of Exports

The blue graph is the cyclical part of the log of real exports of goods and services. The red graph is the cyclical part of the log of real GDP. Real exports are proportionately much more variable than real GDP. Real exports are weakly procyclical—the correlation with the cyclical part of real GDP is 0.38.

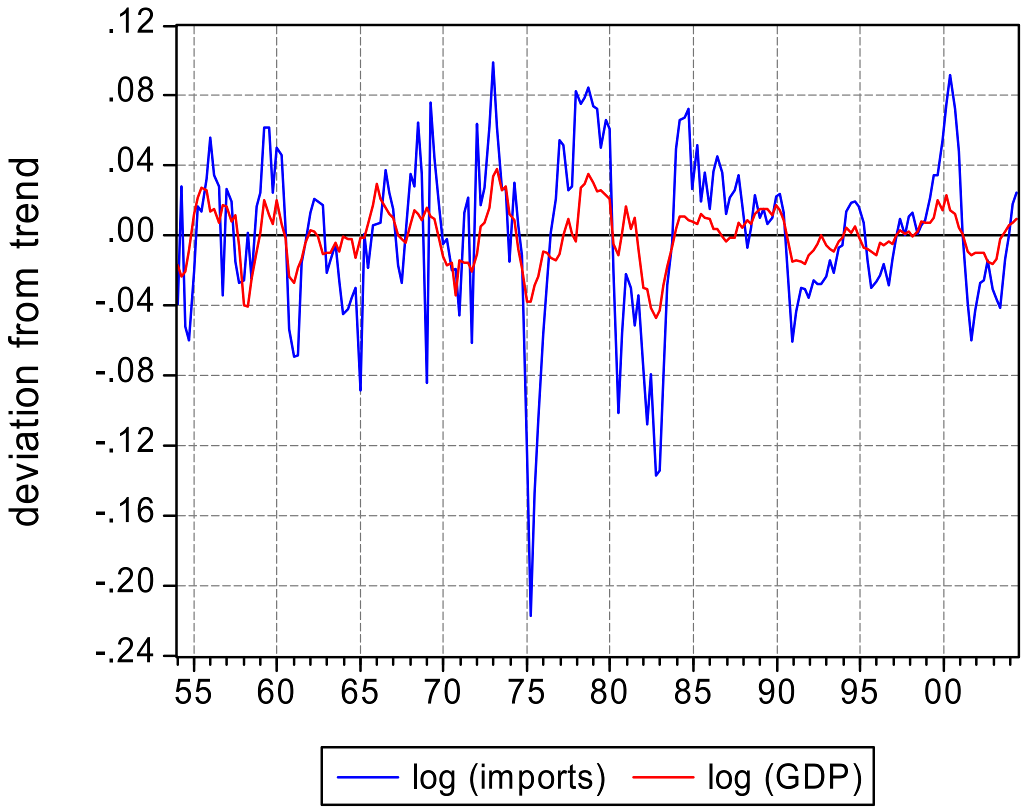


Figure 17.6

Cyclical Behavior of Imports

The blue graph is the cyclical part of the log of real imports of goods and services. The red graph is the cyclical part of the log of real GDP. Real imports are proportionately much more variable than real GDP. Real imports are procyclical—the correlation with the cyclical part of real GDP is 0.71.

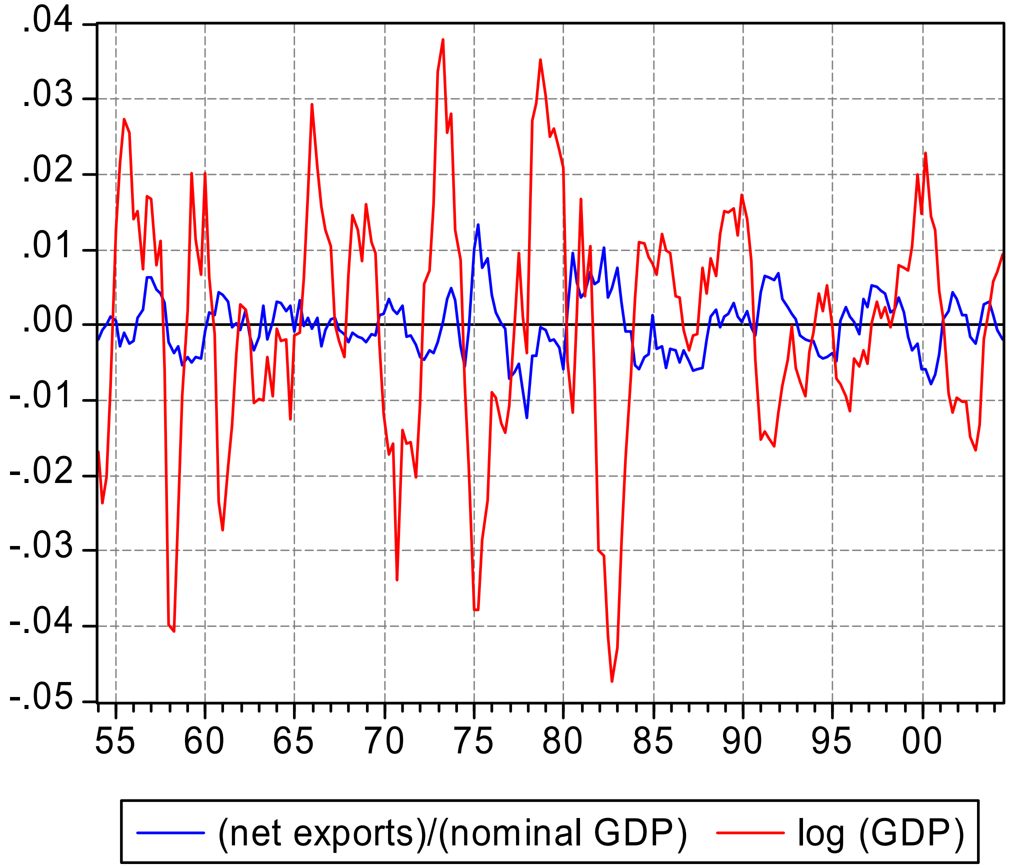


Figure 17.7

Cyclical Behavior of the Trade Balance

The blue graph is the cyclical part of the ratio of net nominal exports (the trade balance) to nominal GDP. The red graph is the cyclical part of the log of real GDP. The ratio of net exports to GDP is weakly countercyclical—the correlation with the cyclical part of real GDP is -0.34.

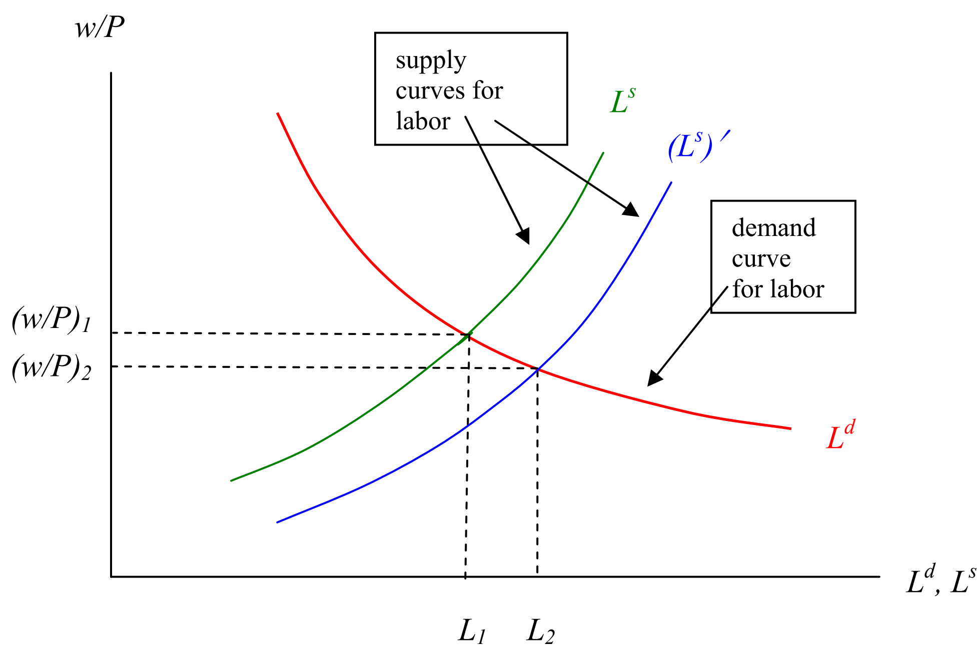


Figure 17.8

Effect of an Improvement in the Terms of Trade on the Labor Market

An increase in the terms of trade, $P/\overline{P}$, does not change the quantity of labor demanded, L^d along the red curve, at a given real wage rate, w/P. However, the rise in $P/\overline{P}$ means that, for a given w/P, the real wage rate as viewed by workers, $w/\overline{P}$, increases. The substitution effect from the higher $w/\overline{P}$ motivates workers to raise the quantity of labor supplied, L^s, at a given w/P. We assume in the graph that this effect dominates over the income effect, which motivates a decline in L^s. Therefore, a rise in $P/\overline{P}$ shifts the labor supply curve rightward from the one labeled L^s (the green curve) to the one labeled $(L^s)'$ (the blue curve). The clearing of the labor market implies that L rises and w/P falls.

January 12, 2005

Chapter 18

Exchange Rates

In the previous chapter we discussed international markets for goods and credit but said nothing about exchange rates. We could not discuss exchange rates because we assumed that all countries used a common currency, such as the U.S. dollar or the euro, and that all prices were quoted in units of this currency. To analyze exchange rates, we have to introduce different types of currency—dollars, euros, pounds, yen, etc.—and allow for prices to be quoted in these different currency units. This chapter makes the necessary extensions to consider these matters.

I. Different Monies and Exchange Rates

Suppose, as in the previous chapter, that the world's quantity of international currency is fixed at the amount $\overline{H}$. We still assume that this currency is denominated in a nominal unit, such as the U.S. dollar, and that the nominal interest rate on international currency is zero.

Assume that each country issues and uses its own currency, instead of using a common currency. Let M^i be the nominal quantity of domestic currency in country i. We measure M^i in domestic currency units, such as dollars, euros, pounds, and yen.

A typical setup is that the central bank of country i holds international currency, H^i, and then issues the domestic currency, M^i. Thus, we would have the simplified balance sheet of a central bank as shown in Table 18.1. The central bank's assets include international currency, H^i, foreign interest-bearing assets, and domestic interest-bearing assets, including bonds issued by the home government. The holdings of domestic earning assets are called the central bank's **domestic credit**. The central bank's liabilities consist of domestic currency, M^i. More realistically, the liabilities would also include deposits held at the central bank by depository institutions and governments.

The domestic price level in country *i*, denoted P^i, gives the number of local currency units, say dollars, that exchange for a unit of goods. To start, think again of a case in which the goods produced in all countries are physically identical. Then the same good sells for P^1 units of the first currency (say dollars) in country 1, P^2 units of the second currency (say euros) in country 2, and so on.

We now introduce a new market, called the **exchange market**, on which participants trade the currency of one country for that of another. For example, traders might exchange euros for U.S. dollars or Japanese yen. The exchange markets establish exchange rates among the various currencies. For convenience, we express all exchange rates as the number of units of domestic currency that trade for $1.00 (U.S.). For example, on November 5, 2004, 0.775 euro exchanged for $1.00, so that each euro was worth $1.29 (where 1.29 = 1/0.775). At the same time, 106 Japanese yen exchanged for $1.00, so that each yen was worth 0.94 cents.

The exchange rates for the yen and the euro with the U.S. dollar determine the exchange rate between the yen and the euro. That is, 0.775 euros buys $1.00, which can then be converted into 106 yen. Thus, the exchange rate between the yen and the euro is

137 (106/0.775) yen per euro. In practice, traders can make these exchanges directly rather than going through U.S. dollars.

Let ε^i (the Greek letter *epsilon*) be the exchange rate for country *i*: ε^i units of country *i*'s currency (say 0.775 euros) exchange for \$1.00. Alternatively, we see that the dollar value of one unit of country *i*'s currency (1 euro) is $1/\varepsilon^i$ (1/0.775 = \$1.29):

ε^i units of currency i buys \$1.00,

1 unit of currency i buys \$$(1/\varepsilon^i)$.

Notice that a *higher* value of the exchange rate, ε^i, means that country *i*'s currency is *less* valuable in terms of dollars because it takes more of country *i*'s currency to buy \$1.00.

For any two countries, *i* and *j*, we observe the exchange rates ε^i and ε^j. These rates prescribe the number of units of each currency that trade for \$1.00. Therefore, ε^i units of currency *i* (say 0.775 euros) can buy ε^j units of currency *j* (say 106 yen). The exchange rate between currencies *i* and *j*—the number of units of currency *i* needed to buy one unit of currency *j*—therefore equals $\varepsilon^i/\varepsilon^j$ (0.775/106 = 0.0073 euro per yen). Alternatively, for one unit of currency *i*, a trader gets $\varepsilon^j/\varepsilon^i$ units of currency *j* (106/0.775 = 137 yen per euro).

Figures 18.1 and 18.2 show the exchange rates between six major currencies—those for France, Germany, Italy, Canada, Japan, and the United Kingdom—and the U.S. dollar from 1950 to 2003. The graphs show the proportionate deviation of the exchange rate for each year from the value that prevailed in 1950. For example, in 1950, it took 1.09 Canadian dollars to buy one U.S. dollar, whereas, in 2003, it took 1.40 Canadian dollars to buy one U.S. dollar. Thus, Figure 18.2 shows that the exchange rate for the Canadian dollar rose by 28% from 1950 to 2003. Remember that a rise in the exchange rate signifies that the currency became less valuable—that is, depreciated—in relation to the U.S. dollar.

For the French franc, German mark, and Italian lira, separate exchange rates with the U.S. dollar applied from 1950 to 1998. However, since 1999, these currencies were each worth a fixed number of euros (6.56 francs per euro, 1.96 marks per euro, and 1936 lira per euro). Therefore, Figure 18.1 shows that the U.S. dollar exchange rates for each of these currencies moved together since 1999. In fact, since 2001, the separate monies ceased to exist, and the euro became the official currency in each of these countries.

II. Purchasing-Power Parity

We are now ready to derive the central theoretical proposition of international finance. This result connects the exchange rate between two currencies to the price levels prevailing in the two countries.

A resident of country i can use a unit of local currency to buy P^i units of goods domestically. Alternatively, the person can use the exchange market to trade his or her unit of domestic currency for $\varepsilon^j/\varepsilon^i$ units of currency j. Then, buying in country j at the price P^j, the person gets $(\varepsilon^j/\varepsilon^i)\cdot(1/P^j)$ units of goods.

Recall our assumptions that goods in country i are identical to those in country j and that all goods are tradable—that is, a resident of country i can buy goods in country j without incurring any transaction costs. In this case, we must have in equilibrium that the options to buy goods domestically or abroad result in the same amount of goods. Otherwise, everyone would want to buy all goods in the cheap country and sell all goods in the expensive country. This result is a version of the law of one price, a concept that we used in chapter 17. Formally, we must have for any countries i and j:

$$1/P^i = (\varepsilon^j/\varepsilon^i)\cdot(1/P^j),$$

goods obtained at home = goods obtained abroad.

We can rearrange the terms in the equation to get

Key equation (purchasing-power parity):

exchange rate between currencies i and j = ratio of prices in countries i and j,

$$\varepsilon^i/\varepsilon^j = P^i/P^j. \tag{18.1}$$

The equation between the exchange rate and the ratio of prices of goods is called **purchasing-power parity (PPP)**. The condition ensures that the purchasing power in terms of goods for each currency is the same regardless of where someone uses the currency to buy goods.

Equation (18.1) says that the level of the exchange rate between currencies i and j, $\varepsilon^i/\varepsilon^j$, equals the level of the price ratio, P^i/P^j. If these two levels are always equal, we must also have that the growth rate of $\varepsilon^i/\varepsilon^j$ equals the growth rate of P^i/P^j:

$$\text{growth rate of } \varepsilon^i/\varepsilon^j = \text{growth rate of } P^i/P^j. \tag{18.2}$$

We can relate the growth rate of $\varepsilon^i/\varepsilon^j$—the exchange rate between currencies i and j—to the individual growth rates of ε^i and ε^j (which are exchange rates of each currency with the U.S. dollar). For given ε^j, the growth rate of $\varepsilon^i/\varepsilon^j$ varies one-for-one and in the same direction with the growth rate of ε^i. For given ε^i, the growth rate of $\varepsilon^i/\varepsilon^j$

varies one-for-one but in the opposite direction with the growth rate of ε^j. Therefore, we have[1]

$$growth\ rate\ of\ \varepsilon^i/\varepsilon^j = growth\ rate\ of\ \varepsilon^i - growth\ rate\ of\ \varepsilon^j,$$

$$growth\ rate\ of\ \varepsilon^i/\varepsilon^j = \Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j,$$

where we used the symbol Δ to represent a change over time. For example, $\Delta\varepsilon^i = \varepsilon^i_{t+1} - \varepsilon^i_t$.

Similarly, we can relate the growth rate of the price ratio, P^i/P^j, to the individual growth rates of P^i and P^j:

$$growth\ rate\ of\ P^i/P^j = growth\ rate\ of\ P^i - growth\ rate\ of\ P^j,$$

$$growth\ rate\ of\ P^i/P^j = \Delta P^i/P^i - \Delta P^j/P^j.$$

Recall from chapter 10 that the growth rate of the price level, $\Delta P/P$, is the inflation rate, π. Therefore, we also have

$$growth\ rate\ of\ P^i/P^j = \pi_i - \pi_j.$$

If we substitute our results on growth rates of $\varepsilon^i/\varepsilon^j$ and P^i/P^j into equation (18.2), we get

Key equation (purchasing-power parity, relative form):

growth rate of exchange rate between i and j = difference in inflation rates,

$$\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j = \pi_i - \pi_j. \tag{18.3}$$

This equation is called the **relative form of PPP**. In contrast, the relation given in levels in equation (18.1) is called the **absolute form of PPP**. The relative form involves growth

[1] The easiest way to prove this result is with calculus.

rates of exchange rates and prices, whereas the absolute form involves levels of exchange rates and prices.

The absolute and relative forms of PPP in equations (18.1) and (18.3) apply when the price levels, P^i and P^j, refer to the same tradable goods in each country. In practice, we would like to interpret P^i and P^j as price indexes for market baskets of goods produced or consumed in countries i and j. For example, we might use GDP deflators or consumer prices indexes (CPIs) to measure P^i and P^j.

When we think of price levels as broad indexes, the PPP conditions need not hold exactly. One reason is that countries specialize in the production of different tradable goods. Changes in the relative prices of these goods—that is, shifts in the terms of trade—will cause the PPP conditions to fail. Suppose, as an example, that country i's terms of trade improve compared to country j. In this case, the price in dollars of goods produced in country i must rise relative to the price in dollars of goods produced in country j. Hence, if P^i and P^j refer to market baskets of produced goods, P^i/P^j would rise for a given exchange rate, $\varepsilon^i/\varepsilon^j$. That is, the PPP conditions in equations (18.1) and (18.3) would not hold.

Another reason for failure of the PPP conditions is that countries produce and consume non-tradable goods and services, such as local labor and real estate. Because non-tradable goods cannot move from one country to another, the purchasing power of a currency in terms of non-tradable goods may depend on where one buys them. Suppose, as an example, that the U.S. dollar price of non-tradable goods and services produced and consumed in country i is high compared to the prices of tradable goods. In this case, if P^i and P^j refer to market baskets of goods consumed or produced, P^i/P^j must be high for a given exchange rate, $\varepsilon^i/\varepsilon^j$. Again, the PPP conditions in equations (18.1) and (18.3) would not hold.

The nearby box shows how the PPP condition in level form (equation [18.1]) works out for a simple good, McDonald's Big Mac hamburger. The table in the box shows that the PPP condition holds fairly well, in the sense that the number of U.S. dollars required to purchase a Big Mac does not differ wildly from one country to another. However, there are differences in these U.S. dollar prices. One reason for the differences is that the provision of Big Macs requires local labor and real estate, and the U.S. dollar prices of these non-tradable goods and services vary across countries. For example, the U.S. dollar price of a Big Mac is lowest in China and the Philippines, places where U.S. dollar wage rates are particularly low.

Purchasing-power parity for the Big Mac

Since 1986, *The Economist* magazine has explored purchasing-power parity by looking at the price in various countries of a simple product, McDonald's Big Mac hamburger. The good is not completely identical in each country, because its provision requires local labor and real estate. Other relevant differences include tax and regulatory policies. Nevertheless, we would predict that the ratios of prices across countries should relate to the market exchange rates.

The Economist of May 27, 2004 reported that the prices of Big Macs and market exchange rates were as shown in the table. If the PPP condition in equation (18.1) held exactly for Big Macs, the ratio of the local price to the U.S.

dollar price ($2.90 in May 2004) would equal the market exchange rate (local currency per U.S. dollar). The table shows that the price ratios and exchange rates were not the same but were positively correlated.

If a country's price ratio exceeds the exchange rate, Big Macs are more expensive in terms of U.S. dollars in that country than in the United States. Or, to put it the other way, the market exchange rate overvalues the local currency by the percentage shown in the last column of the table. If the price ratio is less than the exchange rate, Big Macs are a bargain in the country, and the market exchange rate undervalues the local currency. The table shows that in May 2004 the cheapest places (with the most undervalued currencies) to buy a Big Mac were China, the Philippines, Malaysia, Russia, and Thailand. The most expensive places (with the most overvalued currencies) were Switzerland, Denmark, and Sweden. Similar patterns emerge if one constructs prices of market baskets of goods, not just Big Macs.

International Pricing of Big Macs in May 2004

Country	**Local price (local currency per Big Mac)**	**(local price)/ (U.S. price)**	**Exchange rate (local currency per U.S. dollar)**	**Over (+) or under (-) valuation of local currency**
United States	2.90	1.0	1.00	0.00

Argentina	4.35	1.5	2.94	-0.49
Australia	3.25	1.1	1.43	-0.22
Brazil	5.39	1.9	3.17	-0.41
Britain	4.47	1.5	1.33	0.16
Canada	3.19	1.1	1.37	-0.20
Chile	1400	483	643	-0.25
China	10.41	3.6	8.26	-0.57
Czech Rep.	56.55	19.5	26.55	-0.27
Denmark	27.75	9.6	6.22	0.54
Egypt	10.01	3.5	6.18	-0.44
Euro area	3.07	1.1	0.94	0.13
Hong Kong	12.01	4.1	7.80	-0.47
Hungary	531	183	211	-0.13
Indonesia	16100	5552	9096	-0.39
Japan	262	90.3	112.39	-0.20
Malaysia	5.05	1.7	3.79	-0.54
Mexico	24.01	8.3	11.54	-0.28
New Zealand	4.35	1.5	1.64	-0.09
Peru	8.99	3.1	3.50	-0.11
Philippines	69.02	23.8	56.11	-0.58
Poland	6.29	2.2	3.86	-0.44
Russia	42.05	14.5	29.00	-0.50
Singapore	3.31	1.1	1.72	-0.34
South Africa	12.41	4.3	6.67	-0.36
South Korea	3199	1103	1176	-0.06
Sweden	29.87	10.3	7.58	0.36
Switzerland	6.29	2.2	1.28	0.69
Taiwan	75.11	25.9	33.53	-0.23
Thailand	58.87	20.3	40.60	-0.50
Turkey	3950000	1362069	1531008	-0.11
Venezuela	4400	1517	2972	-0.49

III. Interest-Rate Parity

Assume now that there is a market-determined nominal interest rate in each country. The nominal rate in country *i*, i^i, will be expressed in its own currency, for example, as euros paid per year per euro lent out today.

Consider the situation from the standpoint of a person who is currently holding one U.S. dollar. Suppose that the person is choosing between holding assets in country *i*—say the euro zone—or country *j*—say Japan.[2] In year *t*, the person can exchange the $1.00 for ε_t^i euros. By lending in the euro zone at the nominal interest rate i^i, the person receives in year *t+1* the quantity $\varepsilon_t^i \cdot (1+i^i)$ of euros. If he or she then converts back to U.S. dollars at year *t+1*'s exchange rate ε_{t+1}^i, the amount of dollars received is $\varepsilon_t^i \cdot (1+i^i)/\varepsilon_{t+1}^i$. Similarly, if the person holds assets in country *j* (Japan), the amount of dollars received in year *t+1* is $\varepsilon_t^j \cdot (1+i^j)/\varepsilon_{t+1}^j$.

The first option—holding euro assets—yields $\varepsilon_t^i \cdot (1+i^i)/\varepsilon_{t+1}^i$ dollars. The second option—holding Japanese assets—yields $\varepsilon_t^j \cdot (1+i^j)/\varepsilon_{t+1}^j$ dollars. If there are no restrictions on using the exchange market and holding foreign assets, the options for any countries *i* and *j* must, in equilibrium, yield the same numberof dollars in year *t+1*. Otherwise, people would lend only in the country with the highest yield and borrow only in the country with the lowest yield. The equilibrium condition is therefore

$$\varepsilon_t^i \cdot (1+i^i)/\varepsilon_{t+1}^i = \varepsilon_t^j \cdot (1+i^j)/\varepsilon_{t+1}^j \ ,$$

yield on holding euro assets = yield on holding yen assets.

We can rearrange the terms in the equation to get

[2] The person might also keep the assets in the United States in U.S. dollars.

$$(1+i^i)/(1+i^j) = (\varepsilon^i_{t+1}/\varepsilon^i_t)/(\varepsilon^j_{t+1}/\varepsilon^j_t).$$

Recall that we previously defined the growth rate of the exchange rate from year t to year $t+1$ for country i as $\Delta\varepsilon^i/\varepsilon^i \equiv (\varepsilon^i_{t+1} - \varepsilon^i_t)/\varepsilon^i_t$. Using this definition, we can substitute $\varepsilon^i_{t+1}/\varepsilon^i_t = 1+\Delta\varepsilon^i/\varepsilon^i$. If we also make the parallel substitution for $\varepsilon^j_{t+1}/\varepsilon^j_t$, the equation becomes

$$(1+i^i)/(1+i^j) = (1+\Delta\varepsilon^i/\varepsilon^i)/(1+\Delta\varepsilon^j/\varepsilon^j).$$

The expression on the left-hand side can be shown to be approximately $1+i^i - i^j$. Similarly, the expression on the right-hand side can be shown to be approximately $1+\Delta\varepsilon^i/\varepsilon^i-\Delta\varepsilon^j/\varepsilon^j$. If we make these substitutions and cancel out the "1's" on each side, we get a result called the **interest-rate parity** condition:[3]

Key equation (interest-rate parity):

interest-rate differential for i and j = growth rate of exchange rate between i and j,

(18.4) $$i^i - i^j = \Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j.$$

To think about the interest-rate parity condition in equation (18.4), imagine that the euro-yen (*i, j*) exchange rate is rising over time—that is, the euro is depreciating relative to the yen at the rate $\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j$. In order for assets denominated in euros and yen (*i* and *j*) to yield the same returns, the nominal interest rate in country *i*, i^i, must exceed the nominal interest rate in country *j*, i^j, by the rate of depreciation of country *i*'s currency relative to country *j*'s, $\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j$. That is, the interest rate differential fully

[3] Although we made some approximations to get this result, the equation is exact when we think in terms of continuous time, rather than discrete periods (years).

offsets the growth rate of the exchange rate. Thereby, the return in dollars is the same whether one holds euro assets or yen assets.

In practice, changes in exchange rates would not be known precisely in advance. In that case, the growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j$, in equation (18.4) would be replaced by the expected value of this growth rate. In other words, the difference in nominal interest rates, $i^i - i^j$, offsets only the expected part of the growth rate of the exchange rate.

A number of real-world considerations prevent interest-rate parity from holding exactly. These considerations include the tax treatment of interest income in different countries, uncertainties about asset returns and exchange-rate movements, and governmental restrictions on currency exchanges and asset flows across international borders. For the main developed countries, departures from interest-rate parity are nonzero but tend to be small.

If purchasing-power parity holds in relative form—as in equation (18.3)—the growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j$, equals the difference in the inflation rates between countries i and j, $\pi_i. - \pi_j$. If we substitute $\pi_i. - \pi_j$ for $\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j$ into the interest-rate parity condition in equation (18.4), we get

$$i^i - i^j = \pi_i. - \pi_j.$$

We can rearrange the terms to get

Key equation (equality of real interest rates across countries):

real interest rate in country i = real interest rate in country j,

(18.5) $$i^i - \pi_i = i^j - \pi_j.$$

Thus, the combination of the interest-rate parity condition (equation [18.4]) with the PPP condition in relative form (equation [18.3]) implies that real interest rates are the same in each country.

To get the result in equation (18.5) we assumed in equation (18.4) that the growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j$, equaled the expected growth rate of the exchange rate. As mentioned, the more general condition is that the interest-rate differential, $i^i - i^j$, offsets the expected growth rate of the exchange rate. The relative PPP condition in equation (18.3) then tells us that the expected growth rate of the exchange rate equals the difference in the expected inflation rates for countries *i* and *j*. Consequently, equation (18.5) involves expected, rather than actual, inflation rates. Thus, the equality applies to expected, rather than actual, real interest rates.

For the main developed countries, expected real interest rates on government securities tend to move together, a finding that supports the interest-rate parity condition in equation (18.5).[4] However, detailed empirical investigations do find some divergences in expected real interest rates.[5]

Thinking about divergences in real interest rates

One way to explain divergences across countries in expected real interest rates is to allow for departures from the PPP conditions in equations (18.1) and (18.3). Recall that these conditions need not hold if the underlying price

[4] See Barro and Sala-i-Martin (1990).
[5] See Cumby and Obstfeld (1984) and Mishkin (1984).

levels, P^i and P^j, refer to market baskets of goods produced or consumed in countries i and j.

Let country i represent a place in which the price ratio P^i/P^j is low compared to the exchange rate, $\varepsilon^i / \varepsilon^i$. Suppose that we think of country j as a group of developed countries, such as the euro zone and the United States. Then China is often offered as an example of country *i.* That is, the U.S. dollar price of a market basket of goods in China is relatively cheap. One reason is that Chinese wage rates, measured in U.S. dollars, are comparatively low. (The low wage rates help to explain the relatively low U.S. dollar price of Big Macs in China, as discussed in an earlier box.)

From the perspective of the absolute PPP condition in equation (18.1), if China is country i and the group of developed countries represents country j, we have

$$(18.6) \qquad P^i/P^j < \varepsilon^i / \varepsilon^j,$$

departure from absolute PPP:

relative price of goods in country i < exchange rate.

That is, absolute PPP does not hold, and the U.S. dollar price of a market basket of goods in country i is comparatively low. Equivalently, we can say that, given the price ratio P^i/P^j, the exchange rate between countries i and j, $\varepsilon^i / \varepsilon^j$, is high: the currency of country i (China) looks undervalued.

If goods and services in China are comparatively cheap, in the sense of the inequality in equation (18.6), the demand for Chinese goods and services would be high. This high demand would likely lead over time to increases in U.S. dollar prices of Chinese goods and services compared to U.S. dollar prices of goods and services in the developed countries. That is, the price ratio P^i/P^j would tend to rise relative to the exchange rate, $\varepsilon^i/\varepsilon^j$. One way that this happens is that U.S. dollar wage rates in a country like China tend to increase as the country develops.

We began with a deviation from absolute PPP in equation (18.6). We then found that the relative PPP condition in equation (18.3) would tend not to hold. More specifically, the relative PPP condition should fail in the direction of inflation in country *i* (China) being higher than in country *j* (the group of developed countries), all compared to the growth rate of the exchange rate:

(18.7) $$\pi_i - \pi_j > \Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j,$$

departure from relative PPP:

relative inflation rate for country i > growth rate of exchange rate.

The idea is that the departure from relative PPP in equation (18.7) corrects over time for the departure from absolute PPP in equation (18.6). Since China's prices are comparatively low, relatively high inflation in China is a way to fix the situation. Notice also from equation (18.7) that one possibility is relatively high inflation in China, $\pi_i > \pi_j$. However, another possibility is a reduction of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$-$\Delta\varepsilon^j/\varepsilon^j < 0$. In other words, the Chinese currency could appreciate compared to the currencies of the developed countries.

Suppose now that we combine the inequality from equation (18.7) with the interest-rate parity condition,

(18.4) $$i^i - i^j = \Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j.$$

In this case, we get

$$i^i - i^j < \pi_i - \pi_j.$$

Therefore, if we rearrange terms, we have

(18.8) $$i^i - \pi_i < i^j - \pi_j,$$

real interest rate in country i < real interest rate in country j.

Since the inflation rate in China (country *i*) is relatively high, the real interest rate in China is comparatively low. Therefore, a departure from the relative PPP condition (as in equation [18.7]) is one way to explain cross-country deviations in real interest rates.

IV. Fixed Exchange Rates

Until the early 1970s and except during major wars, economically advanced countries typically maintained **fixed exchange rates** among their currencies.

Figures 18.1 and 18.2 show that, from 1950 to the early 1970s, the exchange rates between six major currencies and the U.S. dollar moved infrequently and by small amounts compared to what came later. For the six countries considered, the main exceptions to fixed exchange rates in this period were the fluctuations in the Canadian dollar rate until the early 1960s and some realignments in the exchange rates for the French franc, German mark, and British pound.

In the previous chapter, we assumed an extreme form of fixed exchange rates—an environment where all countries used a common currency. Since there was only one money, the fixity of exchange rates had to hold. Within a country, this arrangement is so common that it is usually taken for granted—for example, Massachusetts and California use the same U.S. dollar and, therefore, maintain a fixed exchange rate. However, until recently, the typical setup for countries was that each one had its own currency. The major exception since 1999-2001 is the euro, which is now used by 12 Western European countries and is likely to expand to cover additional countries in Europe. At earlier times, the main examples of common currencies were small countries that used another country's currency or shared a single currency. For example, Panama and Ecuador use the U.S. dollar; 12 countries in Africa use the CFA franc, which has been linked aside from one devaluation to the French franc; and 7 islands in the Caribbean use the Caribbean dollar, which is linked to the U.S. dollar.

The fixed-exchange-rate regime that applied to most advanced countries from World War II until the early 1970s was called the **Bretton-Woods System**.[6] Under this system, the participating countries established narrow bands within which they pegged

[6] The system was named in honor of the meeting site, Bretton Woods, New Hampshire, where the regime was set up.

the exchange rate, ε^i, between their currency and the U.S. dollar. Country i's central bank stood ready to buy or sell its currency at the rate of ε^i units per U.S. dollar. For example, the German central bank (Bundesbank) provided dollars for marks when people (or, more likely, financial institutions) wanted to reduce their holdings of marks, and the reverse when people wished to increase their holdings of marks. To manage these exchanges, each central bank maintained a stock of assets as international reserves—for example, U.S. currency or gold or, more likely, interest-bearing assets such as U.S. Treasury securities that could be readily converted into U.S. currency. Then the United States stood ready to exchange U.S. dollars for gold (on the request of foreign official institutions) at a fixed price, which happened to be \$35 per ounce. Thus, by maintaining a fixed exchange rate with the U.S. dollar, each country indirectly pegged its currency to gold. (This setup is called a gold-exchange standard.)

Another historical example of a system of fixed exchange rates is the classical gold standard. In this setup, each central bank directly pegs its currency to gold at a fixed rate of exchange. Britain was effectively on the gold standard from the early eighteenth century until World War I, except for a period of suspension from 1797 to 1821 because of the Napoleonic Wars. Britain returned to the gold standard in 1926 but departed again from the system during the Great Depression in 1931. The United States was on the gold standard from 1879 until the trough of the Great Depression in 1933, when the dollar price of gold was raised from \$20.67 to \$35 per ounce. Earlier periods in the United States involved a greater role for silver in the context of a bimetallic standard. From an international perspective, the high point for the gold standard was from 1890 to 1914.

Under a gold standard (or other commodity standard), each central bank pegs the value of its currency in terms of gold (or other commodities). An ounce of gold might, for example, be set at \$20 in New York and £4 in London (roughly the values prevailing in 1914). In this environment, the exchange rate between U.S. dollars and British pounds would have to be close to five dollars per pound. Otherwise (subject to the costs of shipping gold), it would be profitable for people to buy gold in one country and sell it in the other. As with the Bretton Woods System, the classical gold standard would—if adhered to by the participants—maintain fixed exchange rates among the various currencies.

It is possible for countries to maintain fixed exchange rates in a regime that has no role for gold or other commodities. For example, from 1979 to 1992, several Western European countries kept the exchange rates among their currencies fixed within fairly narrow bands. This arrangement was called the Exchange Rate Mechanism (ERM). This system effectively evolved into the euro, which became the common currency of 12 Western European countries over a transition period from 1999 to 2001. Although most countries in Western Europe use the euro, some important exceptions are the United Kingdom, Sweden, Denmark, and Switzerland.

To see the workings of a system with fixed exchange rates, start by letting P represent the U.S. dollar price of goods in the United States. If the absolute PPP condition from equation (18.1) holds, country i's price level is given by

$$P^i = \varepsilon^i \cdot P. \qquad (18.9)$$

(Note in equation (18.1) that the U.S. dollar exchange rate with itself is unity.) If country i's exchange rate with the U.S. dollar, ε^i, is fixed, equation (18.9) implies that country i's price level, P^i, always moves in lockstep with the U.S. price level, P. We also have from the relative PPP condition in equation (18.3) that country i's inflation rate, π_i, equals the U.S. rate, π:

$$\pi_i = \pi.$$

We can generalize the results by introducing deviations from purchasing-power parity (PPP). As mentioned, these deviations can result from changes in the terms of trade and variations in the relative prices of non-traded goods. However, we would retain the basic result: a country cannot choose independently its exchange rate, ε^i, and its price level, P^i. If the PPP conditions tend to hold in the long run, a fixed exchange rate means that a country's price level must maintain a fixed relation to the U.S. price level. Or, to put it alternatively, the country must, in the long run, experience roughly the same inflation rate as the U.S. inflation rate.

Given fixed exchange rates, the interest-rate parity condition in equation (18.4) implies that country i's nominal interest rate, i^i, equals the U.S. rate, i. Thus, under fixed exchange rates, a single nominal interest rate would prevail in the world. However, this result would not hold precisely if we introduce reasons for departures from interest-rate parity. As mentioned, these reasons include differences in tax treatments and in riskiness of asset returns.

A. The quantity of money under fixed exchange rates

For a closed economy in chapter 10, we stressed the relationship between a country's quantity of money, M^i, and its price level, P^i. Yet we determined the price level P^i in equation (18.9) without saying anything about the country's quantity of money, M^i. Now we investigate the relation between domestic prices and money in an open economy under fixed exchange rates.

It is still the case that the residents of country i demand a quantity of real money, M^i/P^i, that depends on the real GDP, Y^i, the world nominal interest rate, i, and other variables. (We assume that residents of country i use and hold only their own currency.) The condition that all domestic money in country i be willingly held is

(18.10) $$M^i = P^i \cdot £(Y^i , i, \ldots).$$

As in chapter 10, the function $£(\cdot)$ determines the real demand for money. This real demand depends positively on Y^i and negatively on i and may also depend on other variables.

If the absolute PPP condition holds, we can substitute for P^i in equation (18.10) from equation (18.9) to get

(18.11) $$M^i = \varepsilon^i \cdot P \cdot £(Y^i , i, \ldots).$$

Recall that the exchange rate, ε^i, is fixed and, therefore, equals a given number. We assume that the U.S. price level, P, is determined independently of conditions in country i. Therefore, for given determinants of real money demand, $£(\cdot)$, equation (18.11) prescribes country i's nominal quantity of domestic money, M^i. Hence, the quantity of domestic money, M^i, cannot be a free element of choice by country i's central bank. If the bank pegs the exchange rate, ε^i, there is a specific quantity of domestic money, M^i, that is consistent with this exchange rate. This quantity is the amount given by equation (18.11).

To understand these results, assume that the domestic price level, P^i, accords initially with the absolute PPP condition in equation (18.9). Assume further that the domestic quantity of money, M^i, equals the amount given in equation (18.11). In this case, the quantity of domestic money equals the quantity demanded.

Now suppose that the central bank increases the quantity of domestic money, M^i, say by an open-market purchase of domestic government bonds. In Table 18.2, we illustrate this case in step 1 by assuming that domestic currency and the central bank's holdings of interest-bearing domestic assets each rise by \$1 million.

Our analysis of a closed economy in chapter 10 suggests that the increase in the quantity of domestic money, M^i, would raise the domestic price level, P^i. But then P^i would exceed the level dictated by the absolute PPP condition in equation (18.9). Hence, for a given exchange rate, ε^i, goods bought in country *i* would become more expensive compared to goods bought elsewhere. In response, households and businesses would move away from buying goods in country *i* and toward buying goods in other countries (or toward goods imported into country *i* from other countries). This response tends to keep the domestic price level, P^i, from rising; that is, the domestic price level tends to be brought back into line with prices prevailing in other countries. However, at this price level, domestic residents would be unwilling to hold the additional \$1 million of domestic money, M^i. That is, M^i would be greater than the quantity demanded. Households and businesses (or, more realistically, financial institutions) would return their excess domestic currency to the central bank to obtain U.S. dollars or other currencies. Note that, if the central bank is fixing the exchange rate, ε^i, it stands ready to make these exchanges at a fixed conversion ratio. Thus, step 2 in Table 18.2 shows that the quantity of domestic currency, M^i, and the central bank's holdings of international currency, H^i, each decline by \$1 million.

Instead of reducing its holdings of international currency, the central bank could sell off foreign interest-bearing assets in Table 18.1 to get the international currency that people were demanding. Thus, the more general point is that the return of domestic currency to the central bank causes the bank to lose international currency or foreign interest-bearing assets.

To complete the story, we have to assess the central bank's reaction to its loss of international currency or foreign interest-bearing assets. As one possibility, the bank could allow the quantity of domestic money, M^i, to decline. Then, as people return domestic currency to the central bank, the quantity of domestic money falls back toward the level consistent with the absolute PPP condition in equation (18.11). In the end, the central bank holds more domestic interest-bearing assets and less international currency or foreign interest-bearing assets. But the quantity of domestic money, M^i, is unchanged overall. This automatic response of domestic money is a key element of the gold standard and other systems of fixed exchange rates. The mechanism means that, as long as the central bank fixes the exchange rate, ε^i, it lacks control over the quantity of domestic money, M^i, and the domestic price level, P^i.

As another possibility, when the automatic mechanism tends to reduce the quantity of domestic money, M^i, the central bank might attempt to offset this tendency. For example, the bank could conduct another round of open-market operations, which raises M^i. When the bank acts this way, economists say that it attempts to **sterilize** the flow of international currency or foreign interest-bearing assets. By sterilization, economists mean that the central bank attempts to insulate the quantity of domestic money, M^i, from changes in the bank's holdings of international assets. Eventually, this policy can lead to a sufficient drain on international currency and foreign interest-bearing assets so that the central bank becomes unwilling or unable to maintain the exchange rate. That is, the central bank may no longer be willing or able to provide U.S. dollars at the fixed rate of ε^i units of domestic currency per dollar. Instead, there may be a **devaluation**, whereby the exchange rises above ε^i units of domestic currency per U.S. dollar. Thus, the tendency of central banks to sterilize the flows of international assets threatens the viability of fixed exchange rates.[7]

We should mention another possible reaction of government policy to the loss of the central bank's international assets. Recall that this drain resulted in the case just considered from the central bank's excessive monetary creation. This policy tended to raise the domestic price level, P^i, above its absolute PPP value from equation (18.9). To counter the drain of international assets, the home government might impose trade restrictions, which artificially raise the cost of foreign goods for domestic residents. Alternatively, the government might subsidize exports to make these goods cheaper for foreigners. The general point is that the government can interfere with free trade across national borders to prevent the absolute PPP condition from holding. Thus, there are two types of ill effects from excessive monetary expansion in a regime of fixed exchange rates. One is the loss of international assets, an outcome that tends eventually to cause a devaluation. Second, to avoid either devaluation or monetary contraction, the home government may interfere with free trade. In fact, the frequency of these interferences during the post-World War II period was a major argument used by opponents of fixed exchange rates (see Milton Friedman [1968a, chapter 9]).

[7] This discussion uses a framework called the **monetary approach to the balance of payments**. This approach was developed by Mundell (1968, part 2; 1971, part 2). The early origins of this theory are in the eighteenth-century writings of David Hume; see Rotwein (1970).

B. World prices under fixed exchange rates

In a system of fixed exchange rates centered on the U.S. dollar, the absolute PPP condition in equation (18.9) determines each country's price level, P^i, as a ratio to the U.S. price level, P. To complete the picture, we have to determine the U.S. price level. This analysis is similar to the determination of the price level in the closed-economy model of chapter 10. In that model, we determined the price level from the equation of the quantity of money to the quantity demanded. In the present context, we determine the U.S. price level, P, and thereby price levels throughout the world from the equation of the quantity of international currency to the quantity demanded.

Suppose that all countries hold their international currency, H^i, in the form of U.S. dollars (as was reasonably accurate under the Bretton Woods System). Then the total demand for U.S. currency includes the holdings of U.S. residents plus the holdings of foreigners in the form of international currency. The total quantity of U.S. currency, M, now represents the amount available in the whole world. In this setting, we can determine the U.S. price level, P, in the manner of our closed-economy analysis. However, what matters is a comparison of M with the nominal quantity of money demanded in the whole world. A higher M means a higher U.S. price level, P, and a correspondingly higher price level, P^i, in each other country. An increase in the real demand for U.S. currency—whether by U.S. residents or foreigners—lowers the U.S. price level, P, and correspondingly reduces the price level, P^i, in each other country.

Under the international regime that prevailed after World War II, there were a number of factors that constrained the growth of the quantity of U.S. money, M. First, if the Federal Reserve pursued a monetary policy that was inconsistent with stabilization of the U.S. price level, P, U.S. currency would become less attractive as an international medium of exchange. That is, other countries would find it less desirable to fix their exchange rates to the U.S. dollar. This consideration constrained U.S. monetary expansion to the extent that the Federal Reserve wished to maintain the role of the U.S. dollar as the centerpiece of the international monetary system.

More importantly, the United States had a commitment to foreign central banks that they could exchange U.S. dollars for gold at the rate of $35 per ounce. When the U.S. price level rose substantially—as it did starting in the late 1960s—it became attractive for foreign central banks to trade their dollars for gold. As it lost more and more gold, the United States found it increasingly difficult to maintain the dollar price of gold. Eventually, the system broke down at the beginning of the 1970s. Notably, President Richard Nixon decided in 1971 that the United States would no longer provide gold to foreign central banks in exchange for U.S. dollars.

C. Devaluation

Return to a country whose exchange rate, ε^i, is fixed to the U.S. dollar. As discussed, a country that typically fixes its exchange rate occasionally faces pressure to change its rate. Consider, for example, a disturbance that tends to raise the domestic price level, P^i, relative to the U.S. price level, P. The disturbance could be a rapid expansion of domestic currency, M^i, or a decrease in the real demand for country i's money, M^i/P^i. The central bank tends in these circumstances to lose international currency or other foreign assets. In response, the bank may raise the exchange rate, ε^i, that is, devalue the domestic currency in terms of the U.S. dollar.

In the opposite situation, a disturbance tends to lower the domestic price level, P^i, relative to the U.S. price level, P. This disturbance could be a contraction of domestic currency, M^i, or an increase in the real demand for country i's money, M^i/P^i. In this situation, the central bank tends to gain international currency or other foreign assets. The bank may react by lowering the exchange rate, ε^i, that is, by appreciating the domestic currency in terms of the U.S. dollar. Economists call this change a **revaluation**.

Notice that the pressures for devaluation and revaluation in systems of fixed exchange rates are not symmetric. Devaluations relate to losses of international currency or other foreign assets. The threat of running out of international assets provides direct pressure for devaluation—that is, the central bank may reach a situation in which it can no longer maintain its commitment to exchange U.S. dollars for domestic currency at the specified, fixed exchange rate. However, in the reverse situation, the central bank accumulates international currency and other foreign assets. In this case, the pressure to revalue is less direct. Mostly the central bank has to determine that holding vast quantities of foreign assets—such as the U.S. Treasury securities amassed by central banks in Japan, China, and other Asian countries in the first half of the 2000s—is a bad deal. Revaluation is a way to counter this build up of foreign assets.

Devaluations typically do not involve long periods during which the central bank gradually loses international currency or other foreign assets. The expectation of an impending devaluation leads to **speculation**—people (or, more importantly, financial institutions) have the incentive to move out of a currency in advance if they think a devaluation is coming. This speculative decline in the demand for domestic money leads to further losses of international assets and, thereby, tends to hasten the devaluation.

Speculation also applies to cases of impending revaluation. However, a willingness of central banks to accumulate very large quantities of international assets means that these accumulations can occur, along with a fixed exchange rate, for long periods.

Figures 18.1 and 18.2 provide examples of devaluations and revaluations during the mainly fixed-exchange-rate period before the early 1970s. France devalued the franc by a total of 40% in 1957-58, Germany revalued the mark by 5% in 1961 and 7% in 1969, and the United Kingdom devalued the pound by 14% in 1967. In addition, several revaluations occurred in 1971-72, just as the Bretton Woods System was passing into history: Japan revalued the yen by 16%, Germany revalued the mark by 13%, and Switzerland revalued the franc by 13%.

The overall history shows many examples of fixed-exchange-rate systems that ended with substantial devaluations. However, aside from the cases already mentioned up to the early 1970s, it is hard to find fixed-rate regimes that ended with substantial revaluations. In recent times, fixed-exchange-rate systems (or approximations to these) that culminated in large devaluations include the United Kingdom (32% in late 1992), Mexico (97% in late-1994/early-1995), South Korea (88% in late 1997), Malaysia (66% in late 1997), Thailand (114% in late 1997), Indonesia (275% in late 1997), Russia (266% in late 1998), Brazil (71% at the start of 1999), and Argentina (280% in early 2002).

Consider now the effects in the model from a devaluation, that is, an increase in the exchange rate, ε^i. As mentioned, a devaluation is typically a symptom of an underlying disturbance, such as excessive expansion of domestic money, M^i. However, we consider here the effects of an exogenous devaluation—that is, an increase in ε^i that is not accompanied by other economic changes.

If the domestic price level, P^i, does not change, an increase in the exchange rate, ε^i, means that goods purchased in country i become cheaper in terms of U.S. dollars. Thus, for a given U.S. price level, P (and for given prices in other countries), the demand for goods sold by country i rises. This increase in demand tends to raise country i's price level, P^i. Hence, devaluation creates inflationary pressure at home. The absolute PPP condition in equation (18.9) indicates that this pressure continues until P^i rises by the same proportion as the increase in ε^i.

The nominal demand for domestic money, M^i, is given from

$$M^i = P^i \cdot £(Y^i, i, \ldots). \quad (18.10)$$

If we treat the increases in ε^i and P^i as one-time events, the demand for M^i increases in the same proportion as the rise in P^i. The rise in M^i can come about in two ways. First, the domestic central bank may create more money through open-market operations or other means. Second, if the central bank does not act, people would bring international currency to the central bank to acquire more domestic currency. As the bank exchanges domestic for international currency, the quantity of money, M^i, rises.

Notice the two-way direction of causation between devaluation and the behavior of domestic prices and money. First, expansionary monetary policy creates pressure for devaluation. In this sense, domestic inflation tends to cause devaluation. Second, devaluation tends to raise domestic prices and money. In this sense, devaluation is itself inflationary.

Thus far, our analysis treated the increases in the exchange rate, ε^i, and the domestic price level, P^i, as one-time events. However, in practice, countries that devalue once tend to devalue again. Therefore, devaluation today can create expectations of further devaluations in the future. Recall that the interest-rate parity condition for countries *i* and *j* is

$$(18.4) \qquad i^i - i^j = \Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j.$$

If we think of country *j* as the United States and set $\Delta\varepsilon^j/\varepsilon^j = 0$ (because the exchange rate of the U.S. dollar with itself never changes), we have

$$(18.12) \qquad i^i - i = \Delta\varepsilon^i/\varepsilon^i,$$

differential interest rate = growth rate of exchange rate,

where *i* is the U.S. nominal interest rate. We observed that a devaluation may lead to the expectation that $\Delta\varepsilon^i/\varepsilon^i > 0$ (because one devaluation is expected to be followed by others). In this case, equation (18.12) implies that country *i*'s nominal interest rate, i^i, rises for a given U.S. rate, *i*.

The demand-for-money function in equation (18.10) assumed that the home country's nominal interest rate, i^i, equaled the U.S. rate, *i*. More generally, we would replace *i* by i^i on the right-hand side to get

$$(18.13) \qquad M^i = P^i \cdot £(Y^i, i^i, \ldots).$$

An increase in i^i—due in our example to speculation on future devaluation—reduces the nominal demand for domestic currency, M^i. This effect offsets the impact of the increase in P^i, which raises the demand for M^i. Consequently, the speculative effect involving future devaluation makes it uncertain whether M^i will rise or fall in response to today's devaluation.

V. Flexible Exchange Rates

The international system of fixed exchange rates anchored on the U.S. dollar broke down in the early 1970s. One reason was the excessive creation of U.S. dollars and the consequent rise in the U.S. price level after the mid-1960s. This inflation made it increasingly difficult for the United States to maintain convertibility of the dollar into gold at the rate of $35 per ounce. President Nixon decided in 1971 to raise the dollar price of gold and to curb flows of gold from the United States to foreign central banks. These actions signaled the end of the Bretton Woods System, whereby currencies were linked to gold through the U.S. dollar.

Since the early 1970s, many countries have allowed their currencies to vary more or less freely to clear the markets for foreign exchange. We see from Figures 18.1 and 18.2 that the exchange rates of six major currencies with the U.S. dollar have fluctuated substantially since the early 1970s. Groups of countries, such as the members of the European Monetary System from 1979 to 1992 and the euro countries since 1999, have maintained fixed exchange rates among their currencies. Nevertheless, an important development since the early 1970s has been the increased reliance on **flexible exchange rates**. To study this system, we have to extend the model to consider the determination of exchange rates in a flexible-rate system.

We still have the absolute PPP condition:

(18.9) $$P^i = \varepsilon^i \cdot P,$$

where ε^i is the exchange rate for country i with the U.S. dollar and P is the U.S. price level. Deviations from absolute PPP can still arise for reasons discussed before. However, the main point is that, even if absolute PPP holds, P^i need not move in lockstep with P because of adjustments of ε^i.

The monetary authority of country i now has the option of using its policy tools to achieve desired paths of money, M^i, and the price level, P^i. Given the path of P^i, country i can allow the exchange rate, ε^i, to adjust (or float) freely to satisfy the absolute PPP condition in equation (18.9). Thus, the important finding is that country i can choose a monetary policy that is entirely independent of that chosen by the United States or other countries. Since the path of P^i can be chosen, the central bank can also determine the country's inflation rate, π_i. Thus, π_i can deviate in a persistent way from the U.S. inflation rate, π.

We still have the interest-rate parity condition:

$$i^i - i = \Delta\varepsilon^i/\varepsilon^i, \tag{18.12}$$

interest-rate differential = growth rate of the exchange rate.

Since the exchange rate, ε^i, is variable, country i's nominal interest rate, i^i, need no longer equal the U.S. rate, i. However, the difference between the two rates, $i^i - i$, equals the growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$. (More generally, the interest-rate differential equals the expected growth rate of the exchange rate.)

Recall that the PPP condition in relative form is

$$\Delta\varepsilon^i/\varepsilon^i - \Delta\varepsilon^j/\varepsilon^j = \pi_i - \pi_j. \tag{18.3}$$

If we again think of country j as the United States, we can write the relative PPP condition as

$$(18.14) \qquad \pi_i - \pi = \Delta\varepsilon^i/\varepsilon^i,$$

inflation-rate differential = growth rate of exchange rate.

Equation (18.14) fits the data well when applied to countries with sustained high inflation rates, π_i. Not surprisingly, these countries maintained flexible exchange rates with the U.S. dollar over most years, including those prior to the early 1970s. With extremely high inflation, any attempt to fix the exchange rate, ε^i, would result in vast and unsustainable departures from the PPP conditions.

Table 18.3 considers nine countries that maintained high inflation rates from the mid 1950s or later through the mid 1990s. The important finding is that the inflation-rate differential, $\pi_i - \pi$, matches up closely with the growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$.

The match between inflation rates and growth rates of exchange rates is not nearly so good for low-inflation countries. For example, we showed the U.S. dollar exchange for six major countries in Figures 18.1 and 18.2. In the period of mainly flexible exchange rates after the early 1970s, the variations in inflation rates do not explain a lot of the short-term fluctuations in exchange rates for these countries. However, over the longer term, the differences in inflation rates do explain a substantial part of the movements in exchange rates. Table 18.4 shows that, from 1971 to 2003, the countries with higher average inflation-rate differentials, $\pi_i - \pi$, tended to have higher average growth rates of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$.

The difference between the inflation rate differential, $\pi_i - \pi$, and the growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$, gives the rate at which goods produced in country *i* become more expensive compared to goods produced in the United States.[8] For example, in

[8] The results apply to goods produced in a country if the inflation rates refer to GDP deflators, as in Table 18.4.

Table 18.4, France had $\pi_i - \pi = 1.3\%$ per year and $\Delta\varepsilon^i/\varepsilon^i = 0.1\%$ per year. Therefore, from 1971 to 2003, goods produced in France became more expensive compared to those produced in the United States at a rate of 1.2% (1.3% – 0.1%) per year.

From 1971 to 2003, the four Western European countries considered in Table 18.4 were similar in the rates at which goods produced became more expensive compared to goods produced in the United States. These rates were 1.2% per year for France, 1.2% for Germany (-1.0% + 2.2%), 1.5% for Italy (4.7% - 3.2%), and 1.7% for the United Kingdom (2.9% - 1.2%). Note that these rates were about the same for Germany, which had a lower inflation rate ($\pi_i = 3.1\%$) than the United States ($\pi = 4.1\%$), as for Italy, which had a much higher inflation rate ($\pi_i = 8.8\%$). The differences in the growth rates of the exchange rates (-2.2% per year for Germany versus 3.2% per year for Italy) offset the differences in the inflation rates. For Japan, the rate at which goods produced became more expensive compared to goods produced in the United States was higher than for the European countries—2.1% (-1.4% +3.5%) per year. For Canada, the rate was small and negative, -0.3% (0.7% - 1.0%) per year.

VI. Summing Up

In the previous chapter, all countries used a common currency. Now we allow for different currencies and, therefore, for exchange rates between the currencies.

The condition for absolute purchasing-power parity (PPP) connects the exchange rate between two countries to the ratio of price levels prevailing in the countries. This condition comes from the rule of one price—the cost of a good is the same regardless of where one buys or sells the good. The relative form of the PPP condition connects the growth rate of the exchange rate between two countries to the differential in the inflation rates between the two countries.

The PPP conditions explain a lot about cross-country behavior of exchange rates and price levels, especially for high-inflation countries. However, the PPP conditions do not hold exactly if we think of price levels as indexes of market baskets of goods produced or consumed. One reason for failure of the PPP conditions is the presence of non-traded goods and services. Another reason, when countries specialize in different tradable goods, is a change in the terms of trade.

The interest-rate parity condition connects the difference in the nominal interest rate between two countries to the growth rate of the exchange rate between the countries. When combined with the relative form of the PPP condition, the interest-rate parity condition implies that real interest rates are the same across countries. The interest-rate parity condition explains a lot about cross-country behavior of interest rates and exchange rates, especially for the main developed countries. However, this condition need not hold exactly if we allow for differences across countries in tax systems, uncertainties about asset returns and exchange-rate movements, and governmental restrictions on currency exchanges and cross-border asset flows.

Fixed exchange rates applied to the main developed countries under the Bretton Woods system, which prevailed from the end of World War II until the early 1970s. Fixed rates also apply to common-currency systems, such as the euro regime adopted in 12 European countries in 1999-2001. Countries linked through a fixed exchange rate cannot carry out independent monetary policies. Inflation rates have to be similar, and quantities of money have to adjust endogenously to the quantities demanded.

A country in a fixed-exchange rate system sometimes tries to have an independent monetary policy. However, such attempts can lead to losses of international reserves and, consequently, to sharp devaluations of the currency. Revaluations (increases in currency values) can also occur in response to large accumulations of international currency.

Under flexible exchange rates, a country can have an independent monetary policy. This policy can be managed to achieve a target inflation rate. The exchange rate then adjusts to be consistent with the inflation rate.

Questions and Problems

Mainly for review

18.1. Explain how the nominal exchange rate differs from the real exchange rate. Which rate is pegged in a system of fixed exchange rates?

18.2. Explain the conditions for absolute and relative purchasing-power parity in equations (18.1) and (18.3).

18.3. Does the central bank have discretion over the quantity of domestic money in a fixed-exchange-rate system? Show how an attempt to carry out an independent monetary policy can lead to devaluation or revaluation. Why might the attempt lead to trade restrictions?

18.4. In a flexible-exchange-rate system, a country that has a persistently high inflation rate, π, will experience regular increases in its exchange rate, ε. Explain why this happens. Why might the government like this system?

18.5. We mentioned as examples of fixed-exchange-rate systems the classical gold standard, the Bretton Woods system, and a setup with a common currency. Explain how each of these systems operates to maintain a fixed exchange rate.

Problems for discussion

18.x. A shift in the demand for money

Consider an increase in the real demand for money, M^d/P, in the home country.

a. Under a fixed exchange rate, what happens to the home country's price level, P, and nominal quantity of money, M?

b. Under a flexible exchange rate, with a fixed M, what happens to the home country's price level, P, and exchange rate, ε?

18.x. Flexible exchange rates and inflation rates

Equation (18.3) relates the difference between growth rates of exchange rates for two countries to the difference in inflation rates. Use the International Monetary Fund's *International Financial Statistics* to calculate growth rates of exchange rates and inflation rates for some countries. (Use countries other than those shown in Table 18.3.) Do the results accord with equation (18.3)?

18.x. President Nixon's departure from the gold standard in 1971

Under the Bretton Woods System, the United States pegged the price of gold at $35 per ounce.

a. Why did trouble about the gold price arise in 1971?

b. Was Nixon right in eliminating the U.S. commitment to buy and sell gold from foreign official institutions at a fixed price? What other alternatives were there? For example, what was the classical prescription of the gold standard? The French suggested a doubling in the price of gold. Would that change have helped?

18.x. Shipping gold under the gold standard

Suppose (using unrealistic numbers) that the price of gold is \$5 per ounce in New York and £1 per ounce in London. Assume initially that gold can be shipped between New York and London at zero cost.

a. Assume that the dollar-pound exchange rate is \$6 per pound. If a person starts with \$1000 in New York, what can the person do to make a profit? If the cost of shipping gold between New York and London is 1% of the amount shipped, how high does the exchange rate have to rise above \$5 per pound to make this action profitable?

b. Go through the same exercise when the exchange rate is \$4 per pound, from the perspective of someone who starts with £200 in London.

c. The results determine a range of exchange rates around \$5 per pound for which it is unprofitable to ship gold in either direction between New York and London. The upper and lower limits of this range are called gold points. If the exchange rate goes beyond these points, it becomes profitable to ship an unlimited amount of gold. Can you show that the potential to ship gold guarantees that the exchange rate will remain within the gold points?

18.x. Futures contracts on foreign exchange

If a person buys a one-month futures contract on the euro, he or she agrees to purchase euros next month at a dollar exchange rate set today. The buyer of this contract goes long on the euro and does well if the euro appreciate (more than the amount expected) over the month. Similarly, the seller of a futures contract agrees to sell euros next month at a dollar exchange rate set today. The seller goes short on the euro and does well if the euro depreciates (more than expected) over the month.

Consider a euro bond with a maturity of one month. This bond sells for a specific number of euros today and will pay out a stated number of euros in one month. How can a person use the currency futures market to guarantee the dollar rate of return from buying the euro bond and holding it for one month?

18.x. Changes in the quantity of international money

Our analysis treated the quantity of international money as a constant, $\overline{H}$. What modifications have to be made to allow for changes over time in $\overline{H}$? In answering, consider the following regimes:

a. International money consists of U.S. currency.

b. International money costs of bookkeeping entries made by an international organization, such as the International Monetary Fund.

c. International money is a commodity, such as gold.

Table 18.1

Simplified Balance Sheet of a Central Bank

Assets	**Liabilities**
International currency, H^i	Domestic currency, M^i
Foreign interest-bearing assets	
Domestic interest-bearing assets (domestic credit)	

Table 18.2

Effect of an Open-Market Operation on the Central Bank's Balance Sheet

	Assets	**Liabilities**
Step 1	Domestic interest-bearing assets: +\$1 million	Domestic currency, M^i: + \$1 million
Step 2	International currency, H^i: - \$1 million	Domestic currency, M^i: - \$1 million

Table 18.3

A Comparison of Inflation Rates with Growth Rates of Exchange Rates in High-Inflation Countries

Country	Inflation-rate differential, $\pi_i - \pi$ (% per year)	Growth rate of exchange rate, $\Delta\varepsilon^i/\varepsilon^i$ (% per year)
Argentina	68	64
Brazil	81	80
Chile	31	36
Colombia	12	15
Iceland	14	15
Indonesia	7	7
Israel	24	24
Peru	48	46
Uruguay	38	37

The inflation rates, $\boldsymbol{\pi_i}$, are calculated from consumer price indexes. The inflation-rate differential, $\boldsymbol{\pi_i - \pi}$, is the difference between π_i and the U.S. inflation rate, π. The growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$, is the average growth rate of country i's exchange rate with the U.S. dollar. Periods are 1955-96, except 1957-96 for Brazil, 1968-96 for Indonesia, and 1960-96 for Uruguay. The data are from International Monetary Fund, *International Financial Statistics*.

Table 18.4

A Comparison of Inflation Rates with Growth Rates of Exchange Rates in Six Major Economies, 1971-2003

Country	Inflation rate, π_i (% per year)	Inflation-rate differential from U.S., $\pi_i - \pi$ (% per year)	Growth rate of exchange rate, $\Delta\varepsilon^i/\varepsilon^i$ (% per year)
U.S.	4.1	0.0	0.0
France	5.4	1.3	0.1
Germany	3.1	-1.0	-2.2
Italy	8.8	4.7	3.2
Japan	2.7	-1.4	-3.5
Canada	4.8	0.7	1.0
U.K.	6.9	2.9	1.2

The inflation rates, $\boldsymbol{\pi_i}$, are calculated from GDP deflators. The inflation-rate differential, $\boldsymbol{\pi_i - \pi}$, is the difference between π_i and the U.S. inflation rate, π. The growth rate of the exchange rate, $\Delta\varepsilon^i/\varepsilon^i$, is the average growth rate of country *i*'s exchange rate with the U.S. dollar. Periods are 1971-2003. The data are from International Monetary Fund, *International Financial Statistics*.

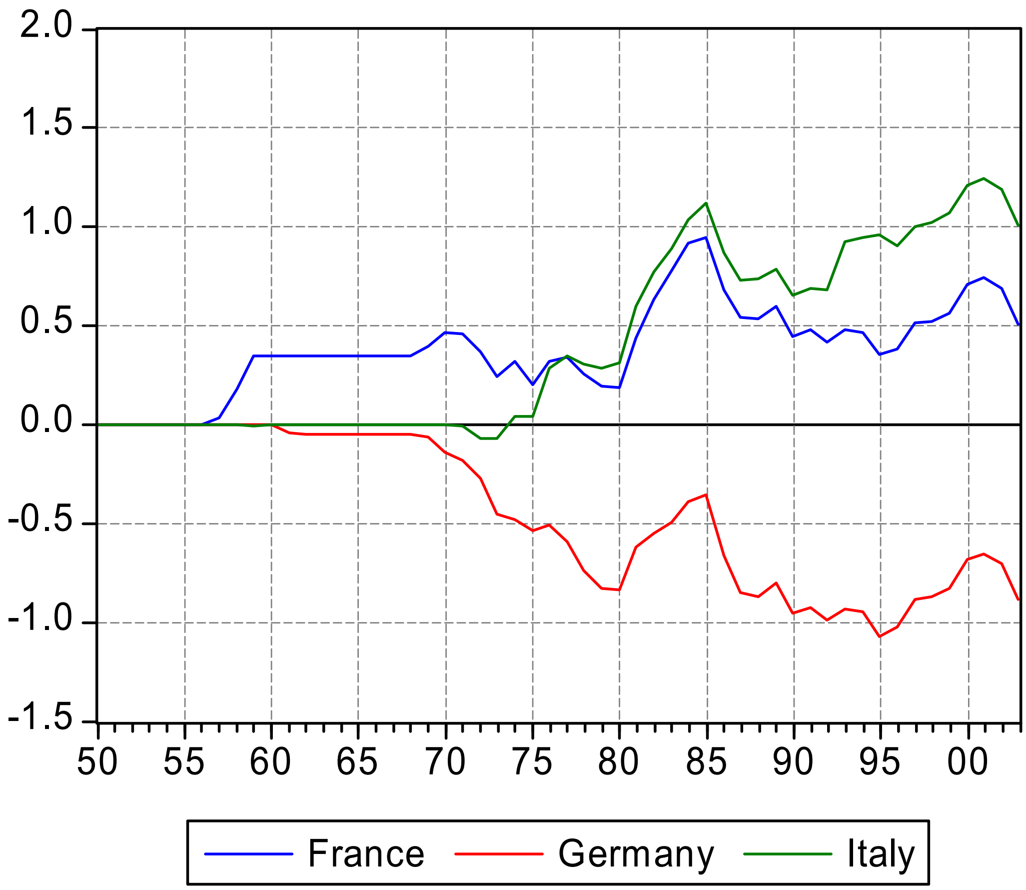


Figure 18.1

Exchange Rates for France, Germany, and Italy

The graphs show the proportionate (logarithmic) deviation of each exchange rate with the U.S. dollar from the value that prevailed in 1950. In 1950, the exchange rates were as follows: France, 3.5 francs per U.S. dollar; Germany, 4.2 marks per U.S. dollar; Italy, 625 lira per U.S. dollar. Since 1999, the exchange rates reflect the movement of the euro against the U.S. dollar. The conversion rates with the euro were fixed as 6.56 French francs, 1.96 German marks, and 1936.3 Italian lira. Data are from International Monetary Fund, *International Financial Statistics*.

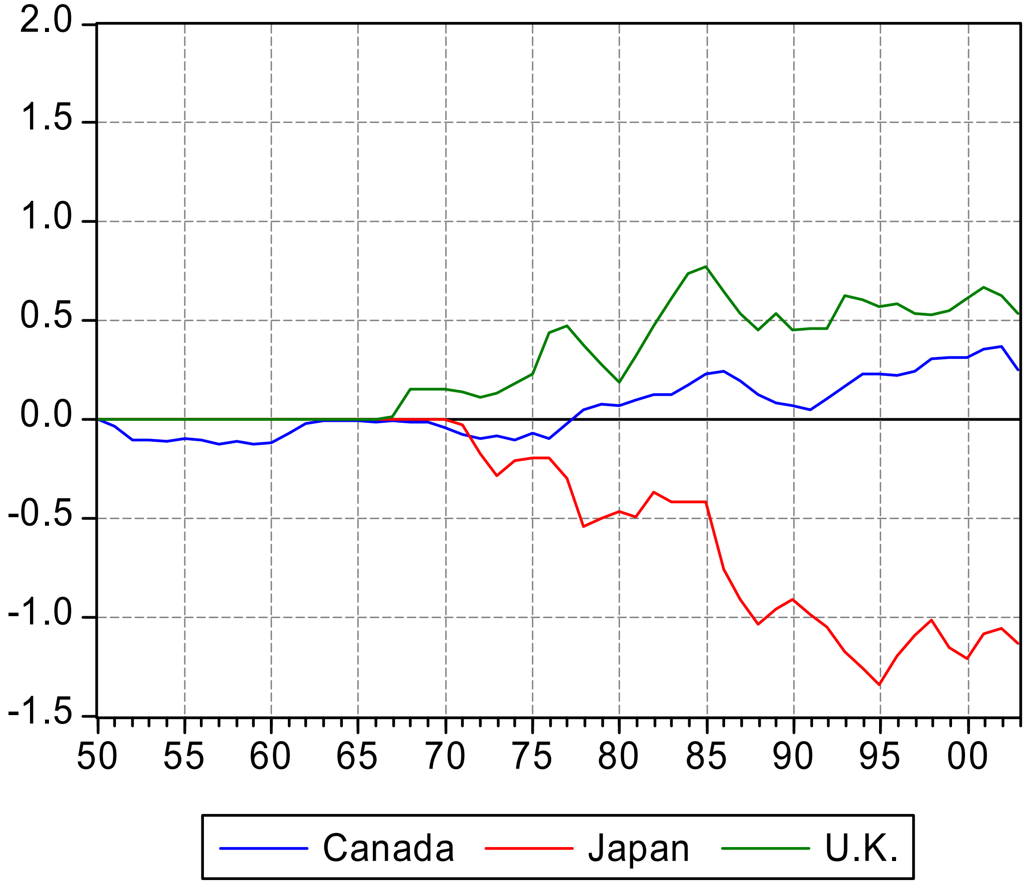


Figure 18.2

Exchange Rates for Canada, Japan, and the United Kingdom

The graphs show the proportionate (logarithmic) deviation of each exchange rate with the U.S. dollar from the value that prevailed in 1950. In 1950, the exchange rates were as follows: Canada, 1.09 Canadian dollars per U.S. dollar; Japan, 361.1 yen per U.S. dollar; United Kingdom, 0.357 pounds per U.S. dollar. For the source, see the notes to Figure 18.1.