The Cost of Production

7

In the last chapter, we examined the firm's production technology—the relationship that shows how factor inputs can be transformed into outputs. Now we will see how the production technology, together with the prices of factor inputs, determines the firm's cost of production.

Given a firm's production technology, managers must decide *how* to produce. As we saw, inputs can be combined in different ways to yield the same amount of output. For example, one can produce a certain output with a lot of labor and very little capital, with very little labor and a lot of capital, or with some other combination of the two. In this chapter we see how the *optimal*—i.e., cost-minimizing—combination of inputs is chosen. We will also see how a firm's costs depend on its rate of output and show how these costs are likely to change over time.

We begin by explaining how *cost* is defined and measured, distinguishing between the concept of cost used by economists, who are concerned about the firm's future performance, and by accountants, who focus on the firm's financial statements. We then examine how the characteristics of the firm's production technology affect costs, both in the short run, when the firm can do little to change its capital stock, and in the long run, when the firm can change all its factor inputs.

We then show how the concept of returns to scale can be generalized to allow for both changes in the mix of inputs and the production of many different outputs. We also show how cost sometimes falls over time as managers and workers learn from experience and make production processes more efficient. Finally, we show how empirical information can be used to estimate cost functions and predict future costs.

CHAPTER OUTLINE

LIST OF EXAMPLES

7.1 MEASURING COST: WHICH COSTS MATTER?

Before we can analyze how firms minimize costs, we must clarify what we mean by *cost* in the first place and how we should measure it. What items, for example, should be included as part of a firm's cost? Cost obviously includes the wages that a firm pays its workers and the rent that it pays for office space. But what if the firm already owns an office building and doesn't have to pay rent? How should we treat money that the firm spent two or three years ago (and can't recover) for equipment or for research and development? We'll answer questions such as these in the context of the economic decisions that managers make.

Economic Cost versus Accounting Cost

Economists think of cost differently from financial accountants, who are usually concerned with keeping track of assets and liabilities and reporting past performance for external use, as in annual reports. Financial accountants tend to take a retrospective view of the firm's finances and operations. As a result **accounting cost**—the cost that financial accountants measure—can include items that an economist would not include and may not include items that economists usually do include. For example, accounting cost includes actual expenses plus depreciation expenses for capital equipment, which are determined on the basis of the allowable tax treatment by the Internal Revenue Service.

• **accounting cost** Actual expenses plus depreciation charges for capital equipment.

Economists—and we hope managers—take a forward-looking view. They are concerned with the allocation of scarce resources. Therefore, they care about what cost is likely to be in the future and about ways in which the firm might be able to rearrange its resources to lower its costs and improve its profitability. As we will see, economists are therefore concerned with **economic cost**, which is the cost of utilizing resources in production. The word *economic* tells us to distinguish between costs that the firm can control and those it cannot. Here the concept of opportunity cost plays an important role.

• **economic cost** Cost to a firm of utilizing economic resources in production, including opportunity cost.

Opportunity Cost

Opportunity cost is the cost associated with opportunities that are forgone by not putting the firm's resources to their best alternative use. For example, consider a firm that owns a building and therefore pays no rent for office space. Does this mean that the cost of office space is zero? While the firm's accountant might say yes, an economist would note that the firm could have earned rent on the office space by leasing it to another company. This forgone rent is the opportunity cost of utilizing the office space and should be included as part of the economic cost of doing business.

• **opportunity cost** Cost associated with opportunities that are forgone when a firm's resources are not put to their best alternative use.

Let's take a look at how opportunity cost can make economic cost differ from accounting cost in the treatment of wages and economic depreciation. Consider an owner who manages her own retail store but chooses not to pay herself a salary. Although no monetary transaction has occurred (and thus no accounting cost is recorded), the business nonetheless incurs an opportunity cost because the owner could have earned a competitive salary by working elsewhere.

Likewise, accountants and economists often treat depreciation differently. When estimating the future profitability of a business, economists and managers are concerned with the capital cost of plant and machinery. This cost involves not only the monetary outlay for buying and then running the machinery, but also the cost associated with wear and tear. When evaluating past performance, cost accountants use tax rules that apply to broadly defined types of assets to determine allowable depreciation in their cost and profit calculations. But these depreciation allowances need not reflect the actual wear and tear on the equipment, which is likely to vary asset by asset.

Sunk Costs

Although an opportunity cost is often hidden, it should be taken into account when making economic decisions. Just the opposite is true of a **sunk cost**: an expenditure that has been made and cannot be recovered. A sunk cost is usually

• **sunk cost** Expenditure that has been made and cannot be recovered.

visible, but after it has been incurred it should always be ignored when making future economic decisions.

Because a sunk cost cannot be recovered, it should not influence the firm's decisions. For example, consider the purchase of specialized equipment for a plant. Suppose the equipment can be used to do only what it was originally designed for and cannot be converted for alternative use. The expenditure on this equipment is a sunk cost. *Because it has no alternative use, its opportunity cost is zero.* Thus it should not be included as part of the firm's economic costs. The decision to buy this equipment may have been good or bad. It doesn't matter. It's water under the bridge and shouldn't affect current decisions.

What if, instead, the equipment could be put to other use or could be sold or rented to another firm? In that case, its use would involve an economic cost—namely, the opportunity cost of using it rather than selling or renting it to another firm.

Now consider a *prospective* sunk cost. Suppose, for example, that the firm has not yet bought the specialized equipment but is merely considering whether to do so. A prospective sunk cost is an *investment*. Here the firm must decide whether that investment in specialized equipment is *economical*—i.e., whether it will lead to a flow of revenues large enough to justify its cost. In Chapter 15, we explain in detail how to make investment decisions of this kind.

As an example, suppose a firm is considering moving its headquarters to a new city. Last year it paid $500,000 for an option to buy a building in the city. The option gives the firm the right to buy the building at a cost of $5,000,000, so that if it ultimately makes the purchase its total expenditure will be $5,500,000. Now it finds that a comparable building has become available in the same city at a price of $5,250,000. Which building should it buy? The answer is the original building. The $500,000 option is a cost that has been sunk and thus should not affect the firm's current decision. What's at issue is spending an additional $5,000,000 or an additional $5,250,000. Because the economic analysis removes the sunk cost of the option from the analysis, the economic cost of the original property is $5,000,000. The newer property, meanwhile, has an economic cost of $5,250,000. Of course, if the new building costs $4,900,000, the firm should buy it and forgo its option.

EXAMPLE 7.1 **Choosing the Location for a New Law School Building**

The Northwestern University Law School has long been located in Chicago, along the shores of Lake Michigan. However, the main campus of the university is located in the suburb of Evanston. In the mid-1970s, the law school began planning the construction of a new building and needed to decide on an appropriate location. Should it be built on the current site, where it would remain near downtown Chicago law firms? Or should it be moved to Evanston, where it would be physically integrated with the rest of the university?

The downtown location had many prominent supporters. They argued in part that it was cost-effective to locate the new building in the city because the university already owned the land. A large parcel of land would have to be purchased in Evanston if the building were to be built there. Does this argument make economic sense?

No. It makes the common mistake of failing to appreciate opportunity costs. From an economic point of view, it is very expensive to locate downtown

because the opportunity cost of the valuable lakeshore location is high: That property could have been sold for enough money to buy the Evanston land with substantial funds left over.

In the end, Northwestern decided to keep the law school in Chicago. This was a costly decision. It may have been appropriate if the Chicago location was particularly valuable to the law school, but it was inappropriate if it was made on the presumption that the downtown land had no cost.

Fixed Costs and Variable Costs

Some costs vary with output, while others remain unchanged as long as the firm is producing any output at all. This distinction will be important when we examine the firm's profit-maximizing choice of output in the next chapter. We therefore divide **total cost (TC or C)**—the total economic cost of production—into two components.

- **total cost (TC or C)** Total economic cost of production, consisting of fixed and variable costs.
- **fixed cost (FC)** Cost that does not vary with the level of output and that can be eliminated only by shutting down.
- **variable cost (VC)** Cost that varies as output varies.

- **Fixed cost (FC):** A cost that does not vary with the level of output and that can be eliminated only by going out of business.
- **Variable cost (VC):** A cost that varies as output varies.

Depending on circumstances, fixed costs may include expenditures for plant maintenance, insurance, heat and electricity, and perhaps a minimal number of employees. They remain the same no matter how much output the firm produces. Variable costs, which include expenditures for wages, salaries, and raw materials used for production, increase as output increases.

Fixed cost does not vary with the level of output—it must be paid even if there is no output. *The only way that a firm can eliminate its fixed costs is by shutting down.*

Shutting Down Shutting down doesn't necessarily mean going out of business. Suppose a clothing company owns several factories, is experiencing declining demand, and wants to reduce output and costs as much as possible at one factory. By reducing the output of that factory to zero, the company could eliminate the costs of raw materials and much of the labor, but it would still incur the fixed costs of paying the factory's managers, security guards, and ongoing maintenance. The only way to eliminate those fixed costs would be to close the doors, turn off the electricity, and perhaps even sell off or scrap the machinery. The company would still remain in business and could operate its remaining factories. It might even be able to re-open the factory it had closed, although doing so could be costly if it involved buying new machinery or refurbishing the old machinery.

Fixed or Variable? How do we know which costs are fixed and which are variable? The answer depends on the time horizon that we are considering. Over a very short time horizon—say, a few months—most costs are fixed. Over such a short period, a firm is usually obligated to pay for contracted shipments of materials and cannot easily lay off workers, no matter how much or how little the firm produces.

On the other hand, over a longer time period—say, two or three years—many costs become variable. Over this time horizon, if the firm wants to reduce its output, it can reduce its workforce, purchase fewer raw materials, and perhaps

even sell off some of its machinery. Over a very long time horizon—say, ten years—nearly all costs are variable. Workers and managers can be laid off (or employment can be reduced by attrition), and much of the machinery can be sold off or not replaced as it becomes obsolete and is scrapped.

Knowing which costs are fixed and which are variable is important for the management of a firm. When a firm plans to increase or decrease its production, it will want to know how that change will affect its costs. Consider, for example, a problem that Delta Air Lines faced. Delta wanted to know how its costs would change if it reduced the number of its scheduled flights by 10 percent. The answer depends on whether we are considering the short run or the long run. Over the short run—say six months—schedules are fixed and it is difficult to lay off or discharge workers. As a result, most of Delta's short-run costs are fixed and won't be reduced significantly with the flight reduction. In the long run—say two years or more—the situation is quite different. Delta has sufficient time to sell or lease planes that are not needed and to discharge unneeded workers. In this case, most of Delta's costs are variable and thus can be reduced significantly if a 10-percent flight reduction is put in place.

Fixed versus Sunk Costs

People often confuse fixed and sunk costs. As we just explained, fixed costs are costs that are paid by a firm that is operating, regardless of the level of output it produces. Such costs can include, for example, the salaries of the key executives and expenses for their office space and support staff, as well as insurance and the costs of plant maintenance. Fixed costs can be avoided if the firm shuts down a plant or goes out of business—the key executives and their support staff, for example, will no longer be needed.

Sunk costs, on the other hand, are costs that have been incurred and *cannot be recovered*. An example is the cost of R&D to a pharmaceutical company to develop and test a new drug and then, if the drug has been proven to be safe and effective, the cost of marketing it. Whether the drug is a success or a failure, these costs cannot be recovered and thus are sunk. Another example is the cost of a chip-fabrication plant to produce microprocessors for use in computers. Because the plant's equipment is too specialized to be of use in any other industry, most if not all of this expenditure is sunk, i.e., cannot be recovered. (Some small part of the cost might be recovered if the equipment is sold for scrap.)

Suppose, on the other hand, that a firm had agreed to make annual payments into an employee retirement plan as long as the firm was in operation, regardless of its output or its profitability. These payments could cease only if the firm went out of business. In this case, the payments should be viewed as a fixed cost.

Why distinguish between fixed and sunk costs? Because fixed costs affect the firm's decisions looking forward, whereas sunk costs do not. Fixed costs that are high relative to revenue and cannot be reduced might lead a firm to shut down—eliminating those fixed costs and earning zero profit might be better than incurring ongoing losses. Incurring a high sunk cost might later turn out to be a bad decision (for example, the unsuccessful development of a new product), but the expenditure is gone and cannot be recovered by shutting down. Of course a *prospective* sunk cost is different and, as we mentioned earlier, would certainly affect the firm's decisions looking forward. (Should the firm, for example, undertake the development of that new product?)

Amortizing Sunk Costs In practice, many firms don't always distinguish between sunk and fixed costs. For example, the semiconductor company that spent $600 million for a chip-fabrication plant (clearly a sunk cost) might **amortize** the expenditure over six years and treat it as a fixed cost of $100 million per year. This is fine as long as the firm's managers understand that shutting down will not make the $100 million annual cost go away. In fact, amortizing capital expenditures this way—spreading them out over many years and treating them as fixed costs—can be a useful way of evaluating the firm's long-term profitability.

• **amortization** Policy of treating a one-time expenditure as an annual cost spread out over some number of years.

Amortizing large capital expenditures and treating them as ongoing fixed costs can also simplify the economic analysis of a firm's operation. As we will see, for example, treating capital expenditures this way can make it easier to understand the tradeoff that a firm faces in its use of labor versus capital. For simplicity, we will usually treat sunk costs in this way as we examine the firm's production decisions. When distinguishing sunk from fixed costs does become essential to the economic analysis, we will let you know.

EXAMPLE 7.2 Sunk, Fixed, and Variable Costs: Computers, Software, and Pizzas

As you progress through this book, you will see that a firm's pricing and production decisions—and its profitability—depend strongly on the structure of its costs. It is therefore important for managers to understand the characteristics of production costs and to be able to identify which costs are fixed, which are variable, and which are sunk. The relative sizes of these different cost components can vary considerably across industries. Good examples include the personal computer industry (where most costs are variable), the computer software industry (where most costs are sunk), and the pizzeria business (where most costs are fixed). Let's look at each of these in turn.

Companies like Dell, Gateway, Hewlett-Packard, and IBM produce millions of personal computers every year. Because computers are very similar, competition is intense, and profitability depends critically on the ability to keep costs down. Most of these costs are variable—they increase in proportion to the number of computers produced each year. Most important is the cost of components: the microprocessor that does much of the actual computation, memory chips, hard disk drives and other storage devices, video and sound cards, etc. Typically, the majority of these components are purchased from outside suppliers in quantities that depend on the number of computers to be produced.

Another important variable cost is labor: Workers are needed to assemble computers and then package and ship them. There is little in the way of sunk costs because factories cost little relative to the value of the company's annual output. Likewise, there is little in the way of fixed costs—perhaps the salaries of the top executives, some security guards, and electricity. Thus, when Dell and Hewlett-Packard think about ways of reducing cost, they focus largely on getting better prices for components or reducing labor requirements—both of which are ways of reducing variable cost.

What about the software programs that run on these personal computers? Microsoft produces the Windows operating system as well as a variety of applications such as Word, Excel, and PowerPoint. But many other firms—some large and some small—also produce software programs that run on personal computers. For such firms, production costs are quite different from those facing

hardware manufacturers. In software production, most costs are *sunk*. Typically, a software firm will spend a large amount of money to develop a new application program. These expenditures cannot be recovered.

Once the program is completed, the company can try to recoup its investment (and make a profit as well) by selling as many copies of the program as possible. The variable cost of producing copies of the program is very small—largely the cost of copying the program to CDs and then packaging and shipping the product. Likewise, the fixed cost of production is small. Because most costs are sunk, entering the software business can involve considerable risk. Until the development money has been spent and the product has been released for sale, an entrepreneur is unlikely to know how many copies can be sold and whether or not he will be able to make money.

Finally, let's turn to your neighborhood pizzeria. For the pizzeria, the largest component of cost is fixed. Sunk costs are fairly low because pizza ovens, chairs, tables, and dishes can be resold if the pizzeria goes out of business. Variable costs are also fairly low—mainly the ingredients for pizza (flour, tomato sauce, cheese, and pepperoni for a typical large pizza might cost $1 or $2) and perhaps wages for a couple of workers to help produce, serve, and deliver pizzas. Most of the cost is fixed—the opportunity cost of the owner's time (he might typically work a 60- or 70-hour week), rent, and utilities. Because of these high fixed costs, most pizzerias (which might charge $12 for a large pizza costing about $3 in variable cost to produce) don't make very high profits.

Marginal and Average Cost

To complete our discussion of costs, we now turn to the distinction between marginal and average cost. In explaining this distinction, we use a specific numerical example of a cost function (the relationship between cost and output) that typifies the cost situation of many firms. The example in shown in Table 7.1. After we explain the concepts of marginal and average cost, we will consider how the analysis of costs differs between the short run and the long run.

Marginal Cost (MC) **Marginal cost**—sometimes called *incremental cost*—is the increase in cost that results from producing one extra unit of output. Because fixed cost does not change as the firm's level of output changes, marginal cost is equal to the increase in variable cost or the increase in total cost that results from an extra unit of output. We can therefore write marginal cost as

• **marginal cost (MC)** Increase in cost resulting from the production of one extra unit of output.

$$MC = \Delta VC/\Delta q = \Delta TC/\Delta q$$

Marginal cost tells us how much it will cost to expand output by one unit. In Table 7.1, marginal cost is calculated from either the variable cost (column 2) or the total cost (column 3). For example, the marginal cost of increasing output from 2 to 3 units is $20 because the variable cost of the firm increases from $78 to $98. (The total cost of production also increases by $20, from $128 to $148. Total cost differs from variable cost only by the fixed cost, which by definition does not change as output changes.)

Average Total Cost (ATC) **Average total cost**, used interchangeably with AC and *average economic cost*, is the firm's total cost divided by its level of output, TC/q. Thus the average total cost of producing at a rate of five units is $36—that is, $180/5. Basically, average total cost tells us the per-unit cost of production.

• **average total cost (ATC)** Firm's total cost divided by its level of output.

TABLE 7.1 A Firm's Costs

Rate of Output (Units per Year)	Fixed Cost (Dollars per Year)	Variable Cost (Dollars per Year)	Total Cost (Dollars per Year)	Marginal Cost (Dollars per Unit)	Average Fixed Cost (Dollars per Unit)	Average Variable Cost (Dollars per Unit)	Average Total Cost (Dollars per Unit)
	(FC) (1)	(VC) (2)	(TC) (3)	(MC) (4)	(AFC) (5)	(AVC) (6)	(ATC) (7)
0	50	0	50	—	—	—	—
1	50	50	100	50	50	50	100
2	50	78	128	28	25	39	64
3	50	98	148	20	16.7	32.7	49.3
4	50	112	162	14	12.5	28	40.5
5	50	130	180	18	10	26	36
6	50	150	200	20	8.3	25	33.3
7	50	175	225	25	7.1	25	32.1
8	50	204	254	29	6.3	25.5	31.8
9	50	242	292	38	5.6	26.9	32.4
10	50	300	350	58	5	30	35
11	50	385	435	85	4.5	35	39.5

• **average fixed cost (AFC)** Fixed cost divided by the level of output.

• **average variable cost (AVC)** Variable cost divided by the level of output.

ATC has two components. **Average fixed cost (AFC)** is the fixed cost (column 1 of Table 7.1) divided by the level of output, FC/q. For example, the average fixed cost of producing 4 units of output is \$12.50 (\$50/4). Because fixed cost is constant, average fixed cost declines as the rate of output increases. **Average variable cost (AVC)** is variable cost divided by the level of output, VC/q. The average variable cost of producing 5 units of output is \$26—that is, \$130/5.

We have now discussed all of the different types of costs that are relevant to production decisions in both competitive and non-competitive markets. Now we turn to how costs differ in the short run versus the long run. This is particularly important for fixed costs. Costs that are fixed in the very short run, e.g., the wages of employees under fixed-term contracts—may not be fixed over a longer time horizon. Similarly, the fixed capital costs of plant and equipment become variable if the time horizon is sufficiently long to allow the firm to purchase new equipment and build a new plant. Fixed costs, however, need not disappear, even in the long run. Suppose, for example, that a firm has been contributing to an employee retirement program. Its obligations, which are fixed in part, may remain even in the long run; they might only disappear if the firm were to declare bankruptcy.

7.2 COST IN THE SHORT RUN

In this section we focus our attention on short-run costs. We turn to long-run costs in Section 7.3.

The Determinants of Short-Run Cost

The data in Table 7.1 show how variable and total costs increase with output in the short run. The rate at which these costs increase depends on the nature of the production process and, in particular, on the extent to which production

involves diminishing marginal returns to variable factors. Recall from Chapter 6 that diminishing marginal returns to labor occur when the marginal product of labor is decreasing. If labor is the only input, what happens as we increase the firm's output? To produce more output, the firm must hire more labor. Then, if the marginal product of labor decreases as the amount of labor hired is increased (owing to diminishing returns), successively greater expenditures must be made to produce output at the higher rate. As a result, variable and total costs increase as the rate of output is increased. On the other hand, if the marginal product of labor decreases only slightly as the amount of labor is increased, costs will not rise so quickly when the rate of output is increased.[1]

In §6.2, we explain that diminishing marginal returns occurs when additional inputs result in decreasing additions to output.

Let's look at the relationship between production and cost in more detail by concentrating on the costs of a firm that can hire as much labor as it wishes at a fixed wage w. Recall that marginal cost MC is the change in variable cost for a 1-unit change in output (i.e., $\Delta VC/\Delta q$). But the change in variable cost is the per-unit cost of the extra labor w times the amount of extra labor needed to produce the extra output ΔL. Because $\Delta VC = w\Delta L$, it follows that

$$MC = \Delta VC/\Delta q = w\Delta L/\Delta q$$

Recall from Chapter 6 that the marginal product of labor MP_L is the change in output resulting from a 1-unit change in labor input, or $\Delta q/\Delta L$. Therefore, the extra labor needed to obtain an extra unit of output is $\Delta L/\Delta q = 1/MP_L$. As a result,

The marginal product of labor is discussed in §6.2.

Marginal product is a MULTIPLIER!

$$MC = w/MP_L \qquad (7.1)$$

Equation (7.1) states that when there is only one variable input, the marginal cost is equal to the price of the input divided by its marginal product. Suppose, for example, that the marginal product of labor is 3 and the wage rate is \$30 per hour. In that case, 1 hour of labor will increase output by 3 units, so that 1 unit of output will require 1/3 additional hour of labor and will cost \$10. The marginal cost of producing that unit of output is \$10, which is equal to the wage, \$30, divided by the marginal product of labor, 3. A low marginal product of labor means that a large amount of additional labor is needed to produce more output—a fact that leads, in turn, to a high marginal cost. Conversely, a high marginal product means that the labor requirement is low, as is the marginal cost. More generally, whenever the marginal product of labor decreases, the marginal cost of production increases, and vice versa.[2]

Diminishing Marginal Returns and Marginal Cost Diminishing marginal returns means that the marginal product of labor declines as the quantity of labor employed increases. As a result, when there are diminishing marginal returns, marginal cost will increase as output increases. This can be seen by looking at the numbers for marginal cost in Table 7.1. For output levels from 0 through 4, marginal cost is declining; for output levels from 4 through 11, however, marginal cost is increasing—a reflection of the presence of diminishing marginal returns.

[1]We are implicitly assuming that because labor is hired in competitive markets, the payment per unit of labor used is the same regardless of the firm's output.

[2]With two or more variable inputs, the relationship is more complex. The basic principle, however, still holds: The greater the productivity of factors, the less the variable cost that the firm must incur to produce any given level of output.

The Shapes of the Cost Curves

Figure 7.1 illustrates how various cost measures change as output changes. The top part of the figure shows total cost and its two components, variable cost and fixed cost; the bottom part shows marginal cost and average costs. These cost curves, which are based on the information in Table 7.1, provide different kinds of information.

Observe in Figure 7.1(a) that fixed cost FC does not vary with output—it is shown as a horizontal line at $50. Variable cost VC is zero when output is zero and then increases continuously as output increases. The total cost curve TC is determined by vertically adding the fixed cost curve to the variable cost curve. Because fixed cost is constant, the vertical distance between the two curves is always $50.

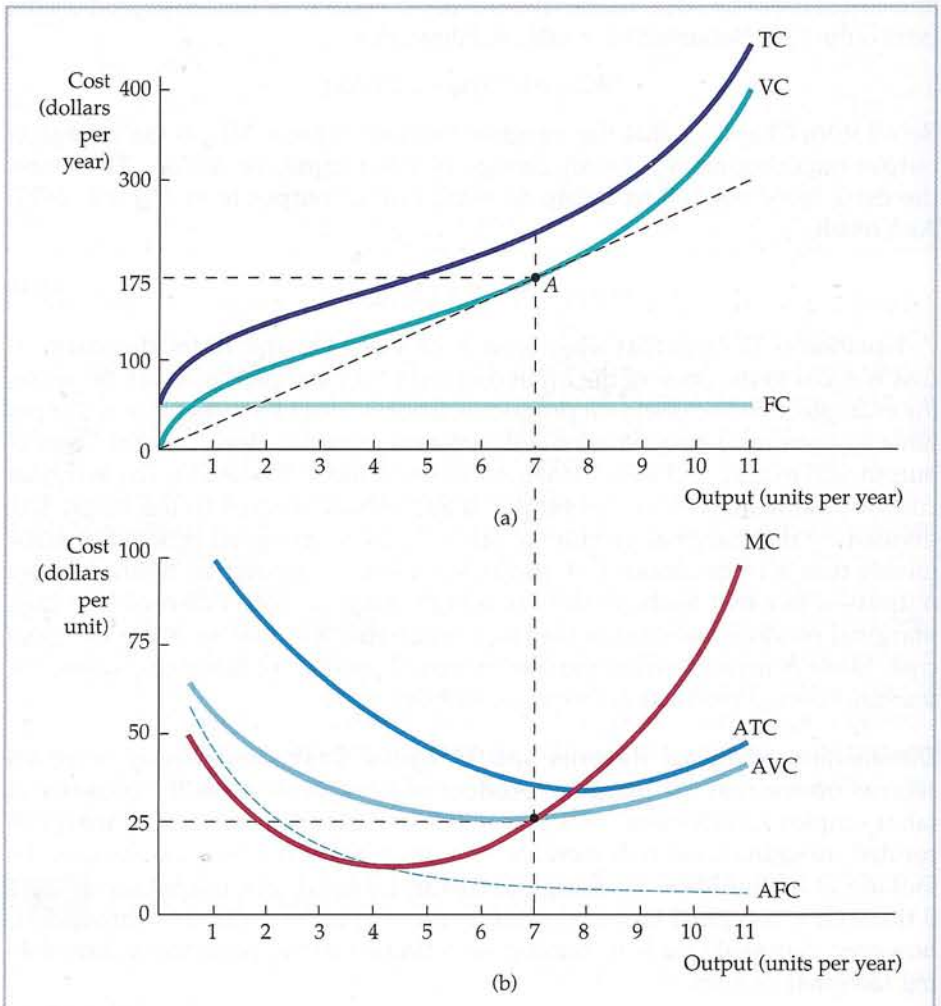

FIGURE 7.1 Cost Curves for a Firm

In **(a)** total cost TC is the vertical sum of fixed cost FC and variable cost VC. In **(b)** average total cost ATC is the sum of average variable cost AVC and average fixed cost AFC. Marginal cost MC crosses the average variable cost and average total cost curves at their minimum points.

Figure 7.1(b) shows the corresponding set of marginal and average variable cost curves.[3] Because total fixed cost is $50, the average fixed cost curve AFC falls continuously from $50 when output is 1, toward zero for large output. The shapes of the remaining curves are determined by the relationship between the marginal and average cost curves. Whenever marginal cost lies below average cost, the average cost curve falls. Whenever marginal cost lies above average cost, the average cost curve rises. When average cost is at a minimum, marginal cost equals average cost.

The Average-Marginal Relationship Marginal and average costs are another example of the average-marginal relationship described in Chapter 6 (with respect to marginal and average product). At an output of 5 in Table 7.1, for example, the marginal cost of $18 is below the average variable cost of $26; thus the average is lowered in response to increases in output. But when marginal cost is $29, which is greater than average variable cost ($25.5), the average increases as output increases. Finally, when marginal cost ($25) and average variable cost ($25) are nearly the same, average variable cost increases only slightly.

The ATC curve shows the average total cost of production. Because average total cost is the sum of average variable cost and average fixed cost and the AFC curve declines everywhere, the vertical distance between the ATC and AVC curves decreases as output increases. The AVC cost curve reaches its minimum point at a lower output than the ATC curve. This follows because MC = AVC at its minimum point and MC = ATC at its minimum point. Because ATC is always greater than AVC and the marginal cost curve MC is rising, the minimum point of the ATC curve must lie above and to the right of the minimum point of the AVC curve.

Another way to see the relationship between the total cost curves and the average and marginal cost curves is to consider the line drawn from origin to point *A* in Figure 7.1(a). In that figure, the slope of the line measures average variable cost (a total cost of $175 divided by an output of 7, or a cost per unit of $25). Because the slope of the VC curve is the marginal cost (it measures the change in variable cost as output increases by 1 unit), the tangent to the VC curve at *A* is the marginal cost of production when output is 7. At *A*, this marginal cost of $25 is equal to the average variable cost of $25 because average variable cost is minimized at this output.

Total Cost as a Flow Note that the firm's output is measured as a flow: The firm produces a certain number of units *per year*. Thus its total cost is a flow—for example, some number of dollars per year. (Average and marginal costs, however, are measured in dollars *per unit*.) For simplicity, we will often drop the time reference, and refer to total cost in dollars and output in units. But you should remember that a firm's production of output and expenditure of cost occur over some time period. In addition, we will often use *cost* (*C*) to refer to total cost. Likewise, unless noted otherwise, we will use *average cost* (AC) to refer to average total cost.

Marginal and average cost are very important concepts. As we will see in Chapter 8, they enter critically into the firm's choice of output level. Knowledge of short-run costs is particularly important for firms that operate in an environment

[3]The curves do not exactly match the numbers in Table 7.1. Because marginal cost represents the change in cost associated with a change in output, we have plotted the MC curve for the first unit of output by setting output equal to $\frac{1}{2}$, for the second unit by setting output equal to $1\frac{1}{2}$, and so on.

in which demand conditions fluctuate considerably. If the firm is currently producing at a level of output at which marginal cost is sharply increasing, and if demand may increase in the future, management might want to expand production capacity to avoid higher costs.

EXAMPLE 7.3 The Short-Run Cost of Aluminum Smelting

Aluminum is a lightweight versatile metal used in a wide variety of applications, including airplanes, automobiles, packaging, and building materials. The production of aluminum begins with the mining of bauxite in such countries as Australia, Brazil, Guinea, Jamaica, and Suriname. Bauxite is an ore that contains a relatively high concentration of alumina (aluminum oxide), which is separated from the bauxite through a chemical refining process. The alumina is then converted to aluminum through a smelting process in which an electric current is used to separate the oxygen atoms from the aluminum oxide molecules. It is this smelting process—which is the most costly step in producing aluminum—that we focus on here.

All of the major aluminum producers, including Alcoa, Alcan, Reynolds, Alumax, and Kaiser, operate smelting plants. A typical smelting plant will have two production lines, each of which produces approximately 300 to 400 tons of aluminum per day. We will examine the short-run cost of production. Thus we consider the cost of operating an existing plant because there is insufficient time in the short run to build additional plants. (It takes about four years to plan, build, and fully equip an aluminum smelting plant.)

Although the cost of a smelting plant is substantial (over $1 billion), we will assume that the plant cannot be sold; the expenditure is therefore sunk and can be ignored. Furthermore, because fixed costs, which are largely for administrative expenses, are relatively small, we will ignore them also. Thus we can focus entirely on short-run variable costs. Table 7.2 shows the average (per-ton) production costs for a typical aluminum smelter.[4] The cost numbers apply to a plant that runs two shifts per day to produce 600 tons of aluminum per day. If prices were sufficiently high, the firm could choose to operate the plant on a three-shifts-per-day basis by asking workers to work overtime. However, wage and maintenance costs would likely increase about 50 percent for this third shift because of the need to pay higher overtime wages. We have divided the cost components in Table 7.2 into two groups. The first group includes those costs that would remain the same at any output level; the second includes costs that would increase if output exceeded 600 tons per day.

Note that the largest cost components for an aluminum smelter are electricity and the cost of alumina; together, they represent about 60 percent of total production costs. Because electricity, alumina, and other raw materials are used in direct proportion to the amount of aluminum produced, they represent per-ton production costs that are constant with respect to the level of output. The costs of labor, maintenance, and freight are also proportional to the level of output, but only when the plant operates two shifts per day. To increase output above 600 tons per day, a third shift would be necessary and would result in a 50-percent increase in the per-ton costs of labor, maintenance, and freight.

The short-run marginal cost and average variable cost curves for the smelting plant are shown in Figure 7.2. For an output q up to 600 tons per day, total

[4]This example is based on Kenneth S. Corts, "The Aluminum Industry in 1994," Harvard Business School Case N9-799-129, April 1999.

TABLE 7.2 Production Costs for Aluminum Smelting ($/ton) (based on an output of 600 tons/day)

Per-ton costs that are constant for all output levels	Output ≤ 600 tons/day	Output > 600 tons/day
Electricity	$316	$316
Alumina	369	369
Other raw materials	125	125
Plant power and fuel	10	10
Subtotal	$820	$820
Per-ton costs that increase when output exceeds 600 tons/day		
Labor	$150	$225
Maintenance	120	180
Freight	50	75
Subtotal	$320	$480
Total per-ton production costs	**$1140**	**$1300**

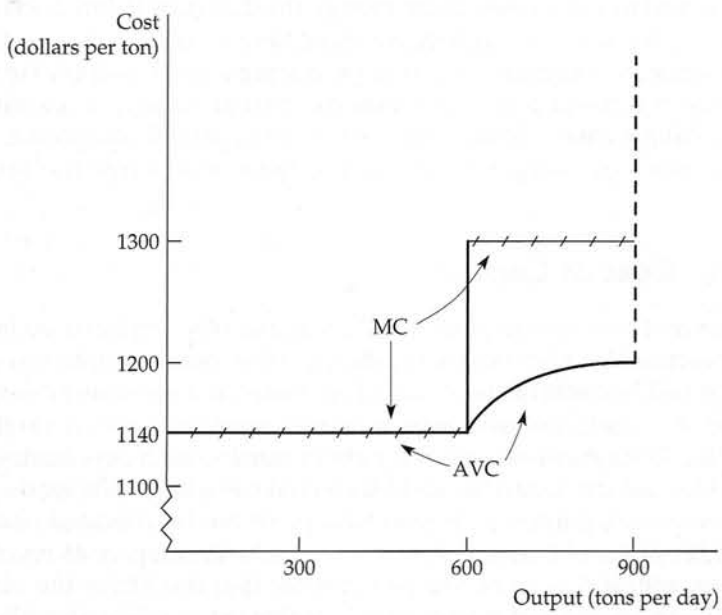

FIGURE 7.2 The Short-Run Variable Costs of Aluminum Smelting

The short-run average variable cost of smelting is constant for output levels using up to two labor shifts. When a third shift is added, marginal cost and average variable cost increase until maximum capacity is reached.

variable cost is $1140q$, so marginal cost and average variable cost are constant at $1140 per ton. If we increase production beyond 600 tons per day by means of a third shift, the marginal cost of labor, maintenance, and freight increases from $320 per ton to $480 per ton, which causes marginal cost as a whole to increase from $1140 per ton to $1300 per ton.

What happens to average variable cost when output is greater than 600 tons per day? When $q > 600$, total variable cost is given by:

$$\text{TVC} = (1140)(600) + 1300(q - 600) = 1300q - 96{,}000$$

Therefore average variable cost is

$$\text{AVC} = 1300 - \frac{96{,}000}{q}$$

As Figure 7.2 shows, when output reaches 900 tons per day, an absolute capacity constraint is reached, at which point the marginal and average costs of production become infinite.

7.3 COST IN THE LONG RUN

In the long run, a firm has much more flexibility. It can expand its capacity by expanding existing factories or building new ones; it can expand or contract its labor force, and in some cases, it can change the design of its products or introduce new products. In this section, we show how a firm can choose its combination of inputs to minimize its cost of producing a given output. We will also examine the relationship between long-run cost and the level of output. We begin by taking a careful look at the cost of using capital equipment. We then show how this cost, along with the cost of labor, enters into the production decision.

The User Cost of Capital

Firms often rent or lease equipment, buildings, and other capital used in the production process. On other occasions, the capital is purchased. In our analysis, however, it will be useful to treat capital as though it were rented even if it was purchased. An illustration will help to explain how and why we do this. Let's suppose that Delta Airlines is thinking about purchasing a new Boeing 777 airplane for $150 million. Even though Delta would pay a large sum for the airplane now, for economic purposes the purchase price can be allocated or *amortized* across the life of the airplane. This will allow Delta to compare its revenues and costs on an *annual flow basis*. We will assume that the life of the airplane is 30 years; the amortized cost is therefore $5 million per year. The $5 million can be viewed as the *annual economic depreciation* for the airplane.

So far, we have ignored the fact that had the firm not purchased the airplane, it could have earned interest on its $150 million. This forgone interest is an *opportunity cost* that must be accounted for. Therefore, the **user cost of capital**—the annual cost of owning and using the airplane instead of selling it or never buying it in the first place—is given by the *sum of the economic depreciation and the*

• **user cost of capital** Annual cost of owning and using a capital asset, equal to economic depreciation plus forgone interest.

interest (i.e., the financial return) that could have been earned had the money been invested elsewhere.[5] Formally,

User Cost of Capital = Economic Depreciation + (Interest Rate)(Value of Capital)

In our example, economic depreciation on the airplane is $5 million per year. Suppose Delta could have earned a return of 10 percent had it invested its money elsewhere. In that case, the user cost of capital is $5 million + (.10)($150 million – depreciation). As the plane depreciates over time, its value declines, as does the opportunity cost of the financial capital that is invested in it. For example, at the time of purchase, looking forward for the first year, the user cost of capital is $5 million + (.10)($150 million) = $20 million. In the tenth year of ownership, the airplane, which will have depreciated by $50 million, will be worth $100 million. At that point, the user cost of capital will be $5 million + (.10)($100 million) = $15 million per year.

We can also express the user cost of capital as a *rate* per dollar of capital:

$$r = \text{Depreciation rate} + \text{Interest rate}$$

For our airplane example, the depreciation rate is 1/30 = 3.33 percent per year. If Delta could have earned a rate of return of 10 percent per year, its user cost of capital would be $r = 3.33 + 10 = 13.33$ percent per year.

As we've already pointed out, in the long run the firm can change all of its inputs. We will now show how the firm chooses the combination of inputs that minimizes the cost of producing a certain output, given information about wages and the user cost of capital. We will then examine the relationship between long-run cost and the level of output.

The Cost-Minimizing Input Choice

We now turn to a fundamental problem that all firms face: *how to select inputs to produce a given output at minimum cost.* For simplicity, we will work with two variable inputs: labor (measured in hours of work per year) and capital (measured in hours of use of machinery per year).

The amount of labor and capital that the firm uses will depend, of course, on the prices of these inputs. We will assume that because there are competitive markets for both inputs, their prices are unaffected by what the firm does. (In Chapter 14 we will examine labor markets that are not competitive.) In this case, the price of labor is simply the *wage rate*, w. But what about the price of capital?

The Price of Capital In the long run, the firm can adjust the amount of capital it uses. Even if the capital includes specialized machinery that has no alternative use, expenditures on this machinery are not yet sunk and must be taken into account; the firm is deciding *prospectively* how much capital to obtain. Unlike labor expenditures, however, large initial expenditures on capital are necessary. In order to compare the firm's expenditure on capital with its ongoing cost of labor, we want to express this capital expenditure as a *flow*—e.g., in dollars per year. To do this, we must amortize the expenditure by spreading it over the lifetime of the capital, and we must also account for the forgone interest that the firm could have earned by investing the money elsewhere. As we have just seen,

[5]More precisely, the financial return should reflect an investment with similar risk. The interest rate, therefore, should include a risk premium. We discuss this point in Chapter 15. Note also that the user cost of capital is not adjusted for taxes; when taxes are taken into account, revenues and costs should be measured on an after-tax basis.

this is exactly what we do when we calculate the *user cost of capital*. As above, the price of capital is its *user cost*, given by r = Depreciation rate + Interest rate.

The Rental Rate of Capital As we noted, capital is often rented rather than purchased. An example is office space in a large office building. In this case, the price of capital is its **rental rate**—i.e., the cost per year for renting a unit of capital.

• **rental rate** Cost per year of renting one unit of capital.

Does this mean that we must distinguish between capital that is rented and capital that is purchased when we determine the price of capital? No. If the capital market is competitive (as we have assumed it is), *the rental rate should be equal to the user cost, r*. Why? Because in a competitive market, firms that own capital (e.g., the owner of the large office building) expect to earn a competitive return when they rent it—namely, the rate of return that they could have earned by investing their money elsewhere, plus an amount to compensate for the depreciation of the capital. *This competitive return is the user cost of capital.*

Many textbooks simply assume that all capital is rented at a rental rate r. As we have just seen, this assumption is reasonable. However, you should now understand *why* it is reasonable: *Capital that is purchased can be treated as though it were rented at a rental rate equal to the user cost of capital.*

For the remainder of this chapter, we will therefore assume that a firm rents all of its capital at a rental rate, or "price," r, just as it hires labor at a wage rate, or "price," w. We will also assume that firms treat any sunk cost of capital as a fixed cost that is spread out over time. We need not, therefore, concern ourselves with sunk costs. Rather, we can now focus on how a firm takes these prices into account when determining how much capital and labor to utilize.[6]

The Isocost Line

• **isocost line** Graph showing all possible combinations of labor and capital that can be purchased for a given total cost.

We begin by looking at the cost of hiring factor inputs, which can be represented by a firm's isocost lines. An **isocost line** shows all possible combinations of labor and capital that can be purchased for a given total cost. To see what an isocost line looks like, recall that the total cost C of producing any particular output is given by the sum of the firm's labor cost wL and its capital cost rK:

$$C = wL + rK \tag{7.2}$$

For each different level of total cost, equation (7.2) describes a different isocost line. In Figure 7.3, for example, the isocost line C_0 describes all possible combinations of labor and capital that cost a total of C_0 to hire.

If we rewrite the total cost equation as an equation for a straight line, we get

$$K = C/r - (w/r)L$$

It follows that the isocost line has a slope of $\Delta K/\Delta L = -(w/r)$, which is the ratio of the wage rate to the rental cost of capital. Note that this slope is similar to the slope of the budget line that the consumer faces (because it is determined solely by the prices of the goods in question, whether inputs or outputs). It tells us that if the firm gave up a unit of labor (and recovered w dollars in cost) to buy w/r units of capital at a cost of r dollars per unit, its total cost of production would remain the same. For example, if the wage rate were \$10 and the rental cost of capital \$5, the firm could replace one unit of labor with two units of capital with no change in total cost.

[6]It is possible, of course, that input prices might increase with demand because of overtime or a relative shortage of capital equipment. We discuss the possibility of a relationship between the price of factor inputs and the quantities demanded by a firm in Chapter 14.

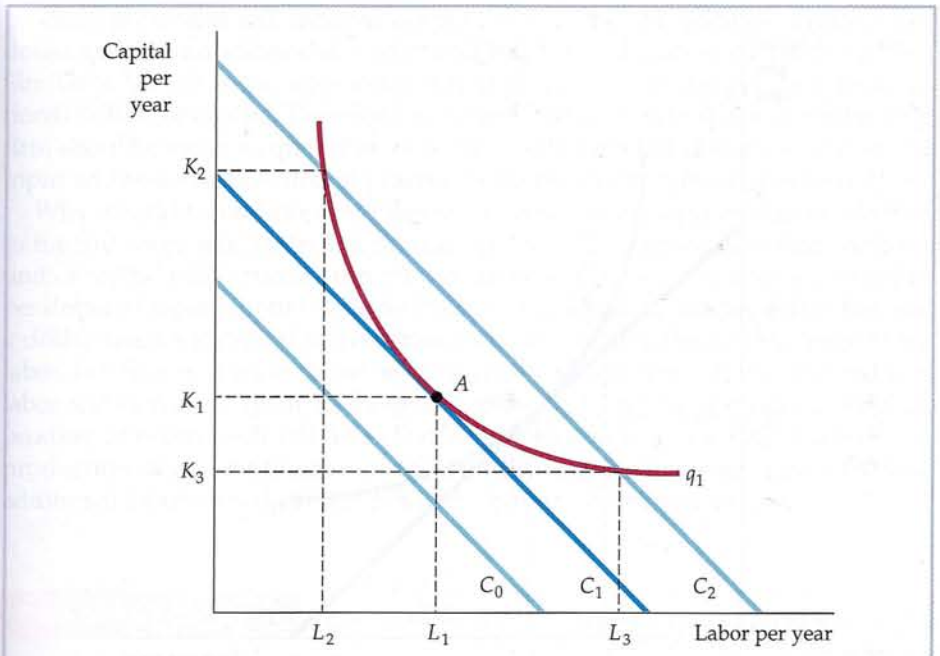

FIGURE 7.3 Producing a Given Output at Minimum Cost

Isocost curves describe the combination of inputs to production that cost the same amount to the firm. Isocost curve C_1 is tangent to isoquant q_1 at A and shows that output q_1 can be produced at minimum cost with labor input L_1 and capital input K_1. Other input combinations—L_2, K_2 and L_3, K_3—yield the same output but at higher cost.

Choosing Inputs

Suppose we wish to produce at an output level q_1. How can we do so at minimum cost? Look at the firm's production isoquant, labeled q_1, in Figure 7.3. The problem is to choose the point on this isoquant that minimizes total cost.

Figure 7.3 illustrates the solution to this problem. Suppose the firm were to spend C_0 on inputs. Unfortunately, no combination of inputs can be purchased for expenditure C_0 that will allow the firm to achieve output q_1. However, output q_1 can be achieved with the expenditure of C_2, either by using K_2 units of capital and L_2 units of labor, or by using K_3 units of capital and L_3 units of labor. But C_2 is not the minimum cost. The same output q_1 can be produced more cheaply, at a cost of C_1, by using K_1 units of capital and L_1 units of labor. In fact, isocost line C_1 is the lowest isocost line that allows output q_1 to be produced. The point of tangency of the isoquant q_1 and the isocost line C_1 at point A gives us the cost-minimizing choice of inputs, L_1 and K_1, which can be read directly from the diagram. At this point, the slopes of the isoquant and the isocost line are just equal.

When the expenditure on all inputs increases, the slope of the isocost line does not change because the prices of the inputs have not changed. The intercept, however, increases. Suppose that the price of one of the inputs, such as labor, were to increase. In that case, the slope of the isocost line $-(w/r)$ would increase in magnitude and the isocost line would become steeper. Figure 7.4 shows this. Initially, the isocost line is C_1, and the firm minimizes its costs of producing output q_1 at A by using L_1 units of labor and K_1 units of capital. When the price of labor increases, the isocost line becomes steeper. The isocost line C_2

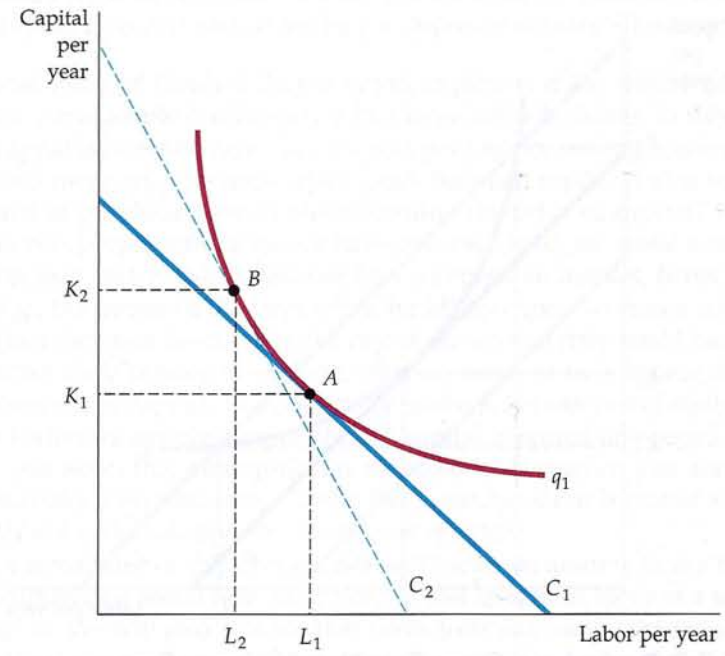

FIGURE 7.4 Input Substitution When an Input Price Changes

Facing an isocost curve C_1, the firm produces output q_1 at point A using L_1 units of labor and K_1 units of capital. When the price of labor increases, the isocost curves become steeper. Output q_1 is now produced at point B on isocost curve C_2 by using L_2 units of labor and K_2 units of capital.

reflects the higher price of labor. Facing this higher price of labor, the firm minimizes its cost of producing output q_1 by producing at B, using L_2 units of labor and K_2 units of capital. The firm has responded to the higher price of labor by substituting capital for labor in the production process.

How does the isocost line relate to the firm's production process? Recall that in our analysis of production technology, we showed that the marginal rate of technical substitution of labor for capital (MRTS) is the negative of the slope of the isoquant and is equal to the ratio of the marginal products of labor and capital:

In §6.3, we explain that the MRTS is the amount by which the input of capital can be reduced when one extra unit of labor is used, so that output remains constant.

$$\text{MRTS} = -\Delta K/\Delta L = \text{MP}_L/\text{MP}_K \quad (7.3)$$

Above, we noted that the isocost line has a slope of $\Delta K/\Delta L = -w/r$. It follows that when a firm minimizes the cost of producing a particular output, the following condition holds:

$$\text{MP}_L/\text{MP}_K = w/r$$

We can rewrite this condition slightly as follows:

$$\text{MP}_L/w = \text{MP}_K/r \quad (7.4)$$

MP_L/w is the additional output that results from spending an additional dollar for labor. Suppose that the wage rate is $10 and that adding a worker to the

production process will increase output by 20 units. The additional output per dollar spent on an additional worker will be 20/10 = 2 units of output per dollar. Similarly, MP_K/r is the additional output that results from spending an additional dollar for capital. Therefore, equation (7.4) tells us that a cost-minimizing firm should choose its quantities of inputs so that the last dollar's worth of any input added to the production process yields the same amount of extra output.

Why must this condition hold for cost minimization? Suppose that in addition to the $10 wage rate, the rental rate on capital is $2. Suppose also that adding a unit of capital will increase output by 20 units. In that case, the additional output per dollar of capital input would be 20/$2 = 10 units of output per dollar. Because a dollar spent for capital is five times more productive than a dollar spent for labor, the firm will want to use more capital and less labor. If the firm reduces labor and increases capital, its marginal product of labor will rise and its marginal product of capital will fall. Eventually, the point will be reached at which the production of an additional unit of output costs the same regardless of which additional input is used. At that point, the firm is minimizing its cost.

EXAMPLE 7.4 The Effect of Effluent Fees on Input Choices

Steel plants are often built on or near rivers. Rivers offer readily available, inexpensive transportation for both the iron ore that goes into the production process and the finished steel itself. Unfortunately, rivers also provide cheap disposal methods for by-products of the production process, called *effluent*. For example, a steel plant processes iron ore for use in blast furnaces by grinding taconite deposits into a fine consistency. During this process, the ore is extracted by a magnetic field as a flow of water and fine ore passes through the plant. One by-product of this process—fine taconite particles—can be dumped in the river at relatively little cost to the firm. Alternative removal methods or private treatment plants are relatively expensive.

Because taconite particles are a nondegradable waste that can harm vegetation and fish, the Environmental Protection Agency (EPA) has imposed an effluent fee—a per-unit fee that the steel firm must pay for the effluent that goes into the river. How should the manager of a steel plant deal with the imposition of this fee to minimize production costs?

Suppose that without regulation the plant is producing 2000 tons of steel per month, using 2000 machine-hours of capital and 10,000 gallons of water (which contains taconite particles when returned to the river). The manager estimates that a machine-hour costs $40 and that dumping each gallon of wastewater in the river costs $10. The total cost of production is therefore $180,000: $80,000 for capital and $100,000 for wastewater. How should the manager respond to an EPA-imposed effluent fee of $10 per gallon of wastewater dumped? The manager knows that there is some flexibility in the production process. If the firm puts into place more expensive effluent treatment equipment, it can achieve the same output with less wastewater.

Figure 7.5 shows the cost-minimizing response. The vertical axis measures the firm's input of capital in machine-hours per month—the horizontal axis measures

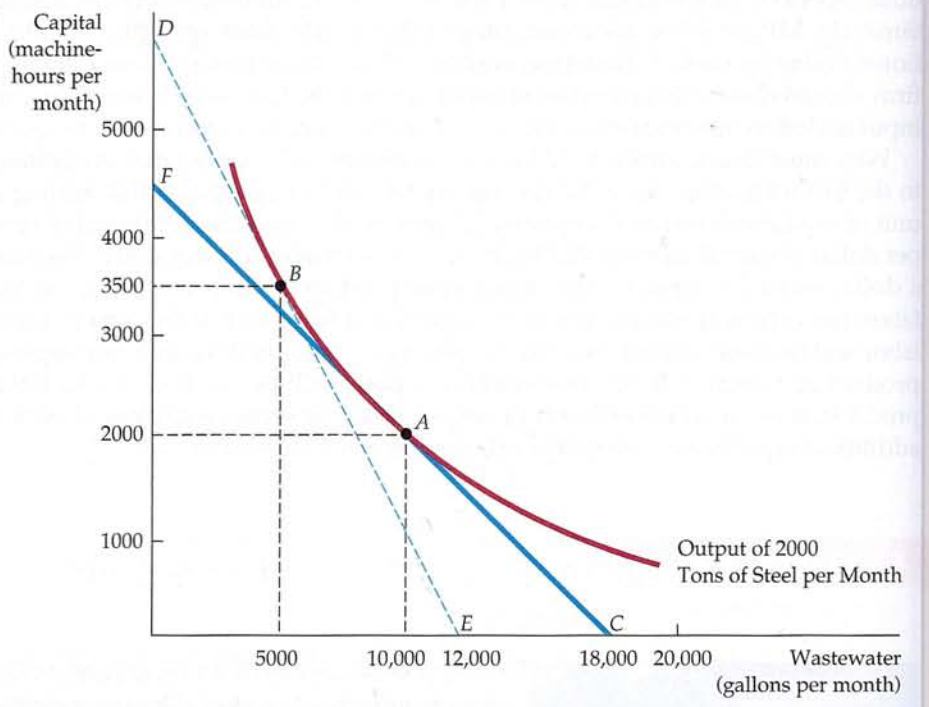

FIGURE 7.5 The Cost-Minimizing Response to an Effluent Fee

When the firm is not charged for dumping its wastewater in a river, it chooses to produce a given output using 10,000 gallons of wastewater and 2000 machine-hours of capital at *A*. However, an effluent fee raises the cost of wastewater, shifts the isocost curve from *FC* to *DE*, and causes the firm to produce at *B*—a process that results in much less effluent.

the quantity of wastewater in gallons per month. First, consider the level at which the firm produces when there is no effluent fee. Point *A* represents the input of capital and the level of wastewater that allows the firm to produce its quota of steel at minimum cost. Because the firm is minimizing cost, *A* lies on the isocost line *FC*, which is tangent to the isoquant. The slope of the isocost line is equal to –\$10/\$40 = –0.25 because a unit of capital costs four times more than a unit of wastewater.

When the effluent fee is imposed, the cost of wastewater increases from \$10 per gallon to \$20: For every gallon of wastewater (which costs \$10), the firm has to pay the government an additional \$10. The effluent fee therefore increases the cost of wastewater relative to capital. To produce the same output at the lowest possible cost, the manager must choose the isocost line with a slope of –\$20/\$40 = –0.5 that is tangent to the isoquant. In Figure 7.5, *DE* is the appropriate isocost line, and *B* gives the appropriate combination of capital and wastewater. The move from *A* to *B* shows that with an effluent fee the use of an alternative production technology that emphasizes the greater use of capital (3500 machine-hours) and less production of wastewater (5000 gallons) is cheaper than the original process which did not emphasize recycling. Note that the total cost of production has increased to \$240,000: \$140,000 for capital, \$50,000 for wastewater, and \$50,000 for the effluent fee.

We can learn two lessons from this decision. First, the more easily factors can be substituted in the production process—that is, the more easily the firm can deal with its taconite particles without using the river for waste treatment—the more effective the fee will be in reducing effluent. Second, the greater the degree of substitution, the less the firm will have to pay. In our example, the fee would have been \$100,000 had the firm not changed its inputs. By moving production from A to B, however, the steel company pays only a \$50,000 fee.

Cost Minimization with Varying Output Levels

In the previous section we saw how a cost-minimizing firm selects a combination of inputs to produce a given level of output. Now we extend this analysis to see how the firm's costs depend on its output level. To do this, we determine the firm's cost-minimizing input quantities for each output level and then calculate the resulting cost.

The cost-minimization exercise yields the result illustrated by Figure 7.6. We have assumed that the firm can hire labor L at w = \$10/hour and rent a unit of capital K for r = \$20/hour. Given these input costs, we have drawn three of the firm's isocost lines. Each isocost line is given by the following equation:

$$C = (\$10/\text{hour})(L) + (\$20/\text{hour})(K)$$

In Figure 7.6(a), the lowest (unlabeled) line represents a cost of \$1000, the middle line \$2000, and the highest line \$3000.

You can see that each of the points A, B, and C in Figure 7.6(a) is a point of tangency between an isocost curve and an isoquant. Point B, for example, shows us that the lowest-cost way to produce 200 units of output is to use 100 units of labor and 50 units of capital; this combination lies on the \$2000 isocost line. Similarly, the lowest-cost way to produce 100 units of output (the lowest unlabeled isoquant) is \$1000 (at point A, $L = 50$, $K = 25$); the least-cost means of getting 300 units of output is \$3000 (at point C, $L = 150$, $K = 75$).

The curve passing through the points of tangency between the firm's isocost lines and its isoquants is its *expansion path*. The **expansion path** describes the combinations of labor and capital that the firm will choose to minimize costs at each output level. As long as the use of both labor and capital increases with output, the curve will be upward sloping. In this particular case we can easily calculate the slope of the line. As output increases from 100 to 200 units, capital increases from 25 to 50 units, while labor increases from 50 to 100 units. For each level of output, the firm uses half as much capital as labor. Therefore, the expansion path is a straight line with a slope equal to

• **expansion path** Curve passing through points of tangency between a firm's isocost lines and its isoquants.

$$\Delta K/\Delta L = (50-25)/(100-50) = \tfrac{1}{2}$$

The Expansion Path and Long-Run Costs

The firm's expansion path contains the same information as its long-run total cost curve, $C(q)$. This can be seen in Figure 7.6(b). To move from the expansion path to the cost curve, we follow three steps:

1. Choose an output level represented by an isoquant in Figure 7.6(a). Then find the point of tangency of that isoquant with an isocost line.

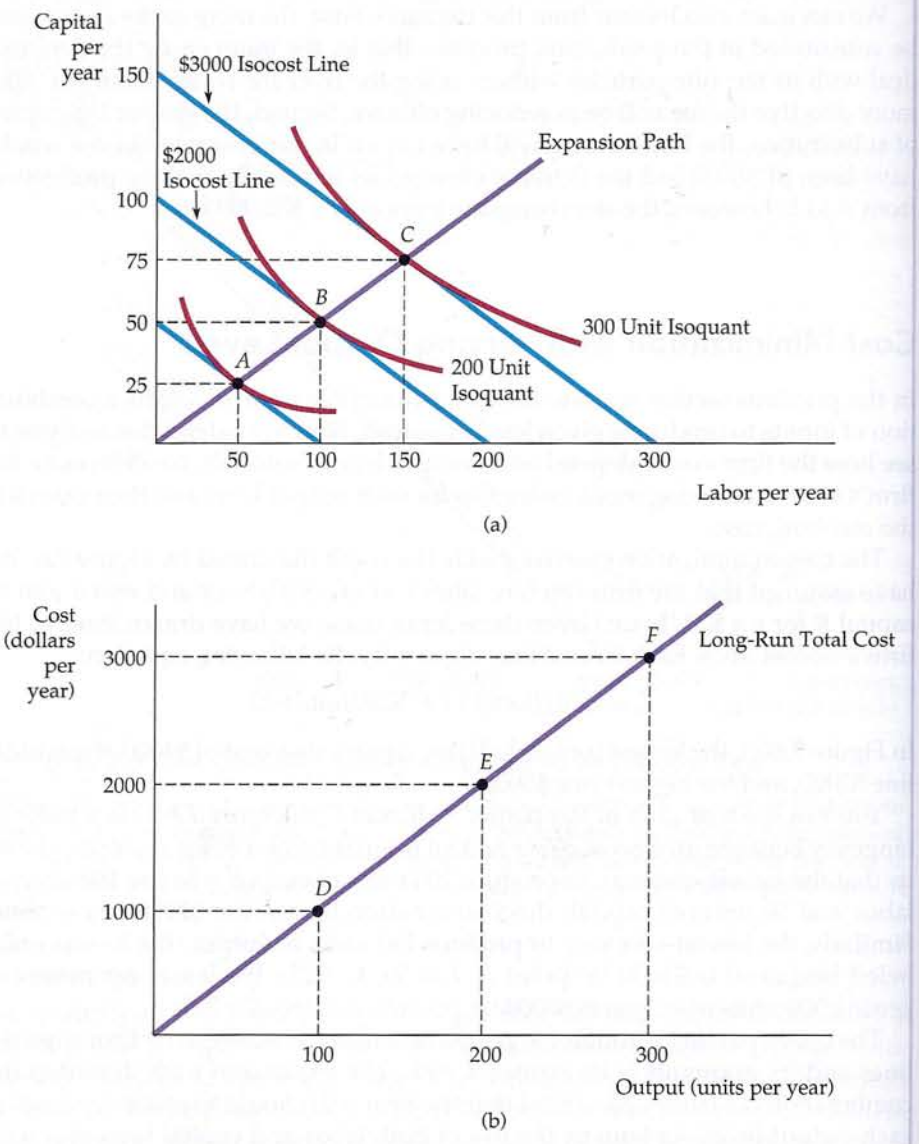

FIGURE 7.6 A Firm's Expansion Path and Long-Run Total Cost Curve

In **(a)**, the expansion path (from the origin through points *A*, *B*, and *C*) illustrates the lowest-cost combinations of labor and capital that can be used to produce each level of output in the long run—i.e., when both inputs to production can be varied. In **(b)**, the corresponding long-run total cost curve (from the origin through points *D*, *E*, and *F*) measures the least cost of producing each level of output.

2. From the chosen isocost line determine the minimum cost of producing the output level that has been selected.
3. Graph the output-cost combination in Figure 7.6(b).

Suppose we begin with an output of 100 units. The point of tangency of the 100-unit isoquant with an isocost line is given by point *A* in Figure 7.6(a). Because *A* lies on the $1000 isocost line, we know that the minimum cost of producing

an output of 100 units in the long run is $1000. We graph this combination of 100 units of output and $1000 cost as point D in Figure 7.6(b). Point D thus represents the $1000 cost of producing 100 units of output. Similarly, point E represents the $2000 cost of producing 200 units which corresponds to point B on the expansion path. Finally, point F represents the $3000 cost of 300 units corresponding to point C. Repeating these steps for every level of output gives the *long-run total cost curve* in Figure 7.6(b)—i.e., the minimum long-run cost of producing each level of output.

In this particular example, the long-run total cost curve is a straight line. Why? Because there are constant returns to scale in production: As inputs increase proportionately, so do outputs. As we will see in the next section, the shape of the expansion path provides information about how costs change with the scale of the firm's operation.

7.4 LONG-RUN VERSUS SHORT-RUN COST CURVES

We saw earlier (see Figure 7.1—page 230) that short-run average cost curves are U-shaped. We will see that long-run average cost curves can also be U-shaped, but different economic factors explain the shapes of these curves. In this section, we discuss long-run average and marginal cost curves and highlight the differences between these curves and their short-run counterparts.

The Inflexibility of Short-Run Production

Recall that we defined the long run as occurring when all inputs to the firm are variable. In the long run, the firm's planning horizon is long enough to allow for a change in plant size. This added flexibility allows the firm to produce at a lower average cost than in the short run. To see why, we might compare the situation in which capital and labor are both flexible to the case in which capital is fixed in the short run.

Figure 7.7 shows the firm's production isoquants. The firm's *long-run expansion path* is the straight line from the origin that corresponds to the expansion path in Figure 7.6. Now, suppose capital is fixed at a level K_1 in the short run. To produce output q_1, the firm would minimize costs by choosing labor equal to L_1, corresponding to the point of tangency with the isocost line AB. The inflexibility appears when the firm decides to increase its output to q_2 without increasing its use of capital. If capital were not fixed, it would produce this output with capital K_2 and labor L_2. Its cost of production would be reflected by isocost line CD.

However, the fact that capital is fixed forces the firm to increase its output by using capital K_1 and labor L_3 at point P. Point P lies on the isocost line EF, which represents a higher cost than isocost line CD. Why is the cost of production higher when capital is fixed? Because the firm is unable to substitute relatively inexpensive capital for more costly labor when it expands production. This inflexibility is reflected in the *short-run expansion path*, which begins as a line from the origin and then becomes a horizontal line when the capital input reaches K_1.

Long-Run Average Cost

In the long run, the ability to change the amount of capital allows the firm to reduce costs. To see how costs vary as the firm moves along its expansion path in the long run, we can look at the long-run average and marginal cost

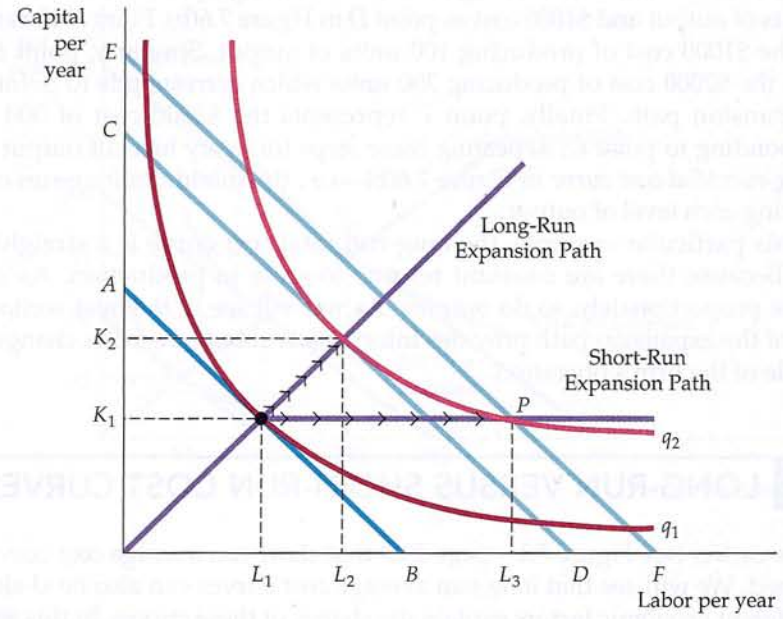

FIGURE 7.7 The Inflexibility of Short-Run Production

When a firm operates in the short run, its cost of production may not be minimized because of inflexibility in the use of capital inputs. Output is initially at level q_1. In the short run, output q_2 can be produced only by increasing labor from L_1 to L_3 because capital is fixed at K_1. In the long run, the same output can be produced more cheaply by increasing labor from L_1 to L_2 and capital from K_1 to K_2.

curves.[7] The most important determinant of the shape of the long-run average and marginal cost curves is the relationship between the scale of the firm's operation and the inputs that are required to minimize its costs. Suppose, for example, that the firm's production process exhibits constant returns to scale at all input levels. In this case, a doubling of inputs leads to a doubling of output. Because input prices remain unchanged as output increases, the average cost of production must be the same for all levels of output.

Suppose instead that the firm's production process is subject to increasing returns to scale: A doubling of inputs leads to more than a doubling of output. In that case, the average cost of production falls with output because a doubling of costs is associated with a more than twofold increase in output. By the same logic, when there are decreasing returns to scale, the average cost of production must be increasing with output.

We saw that the long-run total cost curve associated with the expansion path in Figure 7.6(a) was a straight line from the origin. In this constant-returns-to-scale case, the long-run average cost of production is constant: It is unchanged as output increases. For an output of 100, long-run average cost is \$1000/100 = \$10 per unit. For an output of 200, long-run average cost is \$2000/200 = \$10 per unit;

[7] In the short run, the shapes of the average and marginal cost curves were determined primarily by diminishing returns. As we showed in Chapter 6, diminishing returns to each factor is consistent with constant (or even increasing) returns to scale.

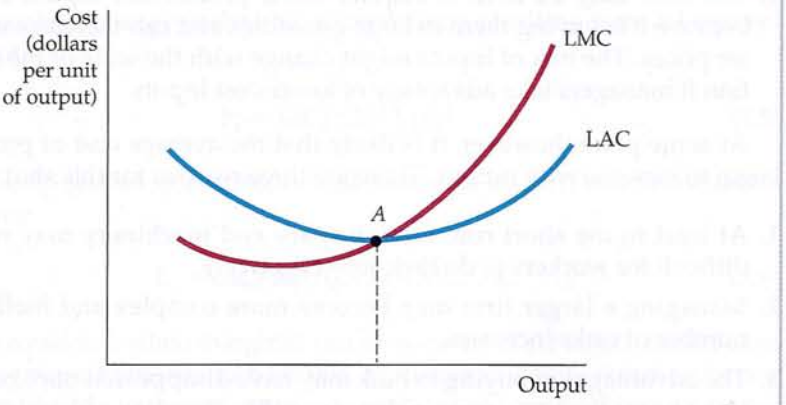

FIGURE 7.8 Long-Run Average and Marginal Cost

When a firm is producing at an output at which the long-run average cost LAC is falling, the long-run marginal cost LMC is less than LAC. Conversely, when LAC is increasing, LMC is greater than LAC. The two curves intersect at *A*, where the LAC curve achieves its minimum.

for an output of 300, average cost is also \$10 per unit. Because a constant average cost means a constant marginal cost, the long-run average and marginal cost curves are given by a horizontal line at a \$10/unit cost.

Recall that in the last chapter we examined a firm's production technology that exhibits first increasing returns to scale, then constant returns to scale, and eventually decreasing returns to scale. Figure 7.8 shows a typical **long-run average cost curve (LAC)** consistent with this description of the production process. Like the **short-run average cost curve (SAC)**, the long-run average cost curve is U-shaped, but the source of the U-shape is increasing and decreasing returns to scale, rather than diminishing returns to a factor of production.

The **long-run marginal cost curve (LMC)** can be determined from the long-run average cost curve; it measures the change in long-run total costs as output is increased incrementally. LMC lies below the long-run average cost curve when LAC is falling and above it when LAC is rising.[8] The two curves intersect at *A*, where the long-run average cost curve achieves its minimum. In the special case in which LAC is constant, LAC and LMC are equal.

- **long-run average cost curve (LAC)** Curve relating average cost of production to output when all inputs, including capital, are variable.
- **short-run average cost curve (SAC)** Curve relating average cost of production to output when level of capital is fixed.
- **long-run marginal cost curve (LMC)** Curve showing the change in long-run total cost as output is increased incrementally by 1 unit.

Economies and Diseconomies of Scale

As output increases, the firm's average cost of producing that output is likely to decline, at least to a point. This can happen for the following reasons:

1. If the firm operates on a larger scale, workers can specialize in the activities at which they are most productive.
2. Scale can provide flexibility. By varying the combination of inputs utilized to produce the firm's output, managers can organize the production process more effectively.

[8]Recall that $AC = TC/q$. It follows that, $\Delta AC/\Delta q = [q(\Delta TC/\Delta q) - TC]/q^2 = (MC - AC)/q$. Clearly, when AC is increasing, $\Delta AC/\Delta q$ is positive and $MC > AC$. Correspondingly, when AC is decreasing, $\Delta AC/\Delta q$ is negative and $MC < AC$.

3. The firm may be able to acquire some production inputs at lower cost because it is buying them in large quantities and can therefore negotiate better prices. The mix of inputs might change with the scale of the firm's operation if managers take advantage of lower-cost inputs.

At some point, however, it is likely that the average cost of production will begin to increase with output. There are three reasons for this shift:

1. At least in the short run, factory space and machinery may make it more difficult for workers to do their jobs effectively.
2. Managing a larger firm may become more complex and inefficient as the number of tasks increases.
3. The advantages of buying in bulk may have disappeared once certain quantities are reached. At some point, available supplies of key inputs may be limited, pushing their costs up.

To analyze the relationship between the scale of the firm's operation and the firm's costs, we need to recognize that when input proportions do change, the firm's expansion path is no longer a straight line, and the concept of returns to scale no longer applies. Rather, we say that a firm enjoys **economies of scale** when it can double its output for less than twice the cost. Correspondingly, there are **diseconomies of scale** when a doubling of output requires more than twice the cost. The term *economies of scale* includes increasing returns to scale as a special case, but it is more general because it reflects input proportions that change as the firm changes its level of production. In this more general setting, a U-shaped long-run average cost curve characterizes the firm facing economies of scale for relatively low output levels and diseconomies of scale for higher levels.

• **economies of scale**
Situation in which output can be doubled for less than a doubling of cost.

• **diseconomies of scale**
Situation in which a doubling of output requires more than a doubling of cost.

In §6.4, we explain that increasing returns to scale occurs when output more than doubles as inputs are doubled proportionately.

To see the difference between returns to scale (in which inputs are used in constant proportions as output is increased) and economies of scale (in which input proportions are variable), consider a dairy farm. Milk production is a function of land, equipment, cows, and feed. A dairy farm with 50 cows will use an input mix weighted toward labor and not equipment (i.e., cows are milked by hand). If all inputs were doubled, a farm with 100 cows could double its milk production. The same will be true for the farm with 200 cows, and so forth. In this case, there are constant returns to scale.

Large dairy farms, however, have the option of using milking machines. If a large farm continues milking cows by hand, regardless of the size of the farm, constant returns would continue to apply. However, when the farm moves from 50 to 100 cows, it switches its technology toward the use of machines, and, in the process, is able to reduce its average cost of milk production from 20 cents per gallon to 15 cents per gallon. In this case, there are economies of scale.

This example illustrates the fact that a firm's production process can exhibit constant returns to scale, but still have economies of scale as well. Of course, firms can enjoy both increasing returns to scale and economies of scale. It is helpful to compare the two:

Increasing Returns to Scale:	Output more than doubles when the quantities of all inputs are doubled.
Economies of Scale:	A doubling of output requires less than a doubling of cost.

Economies of scale are often measured in terms of a cost-output elasticity, E_C. E_C is the percentage change in the cost of production resulting from a 1-percent increase in output:

$$E_C = (\Delta C/C)/(\Delta q/q) \quad \textbf{(7.5)}$$

To see how E_C relates to our traditional measures of cost, rewrite equation (7.5) as follows:

$$E_C = (\Delta C/\Delta q)/(C/q) = \text{MC}/\text{AC} \quad \textbf{(7.6)}$$

Clearly, E_C is equal to 1 when marginal and average costs are equal. In that case, costs increase proportionately with output, and there are neither economies nor diseconomies of scale (constant returns to scale would apply if input proportions were fixed). When there are economies of scale (when costs increase less than proportionately with output), marginal cost is less than average cost (both are declining) and E_C is less than 1. Finally, when there are diseconomies of scale, marginal cost is greater than average cost and E_C is greater than 1.

The Relationship between Short-Run and Long-Run Cost

Figure 7.9 shows the relationship between short-run and long-run cost. Assume that a firm is uncertain about the future demand for its product and is considering three alternative plant sizes. The short-run average cost curves for the three plants are given by SAC_1, SAC_2, and SAC_3. The decision is important because, once built, the firm may not be able to change the plant size for some time.

Figure 7.9 illustrates the case in which there are three possible plant sizes. If the firm expects to produce q_0 units of output, then it should build the smallest plant. Its average cost of production would be \$8. (If it then decided to produce

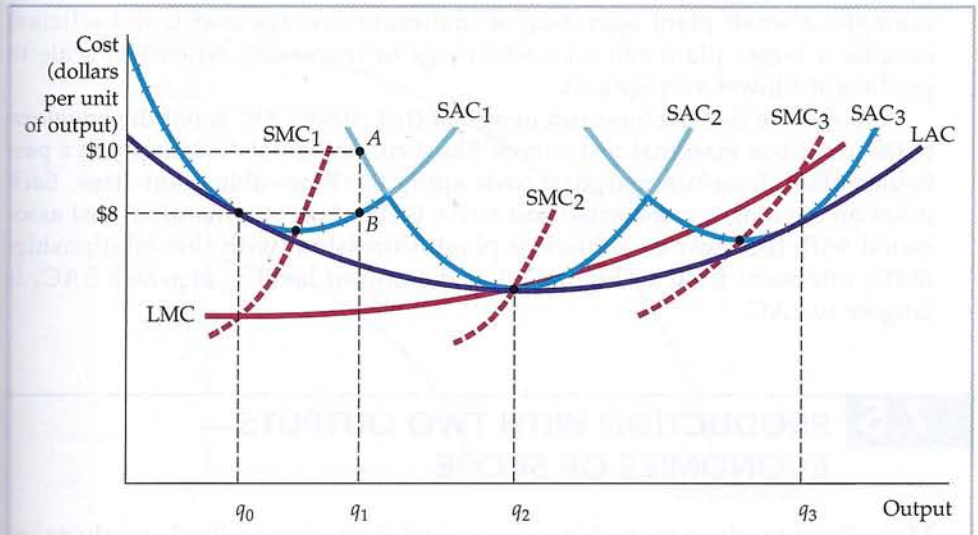

FIGURE 7.9 Long-Run Cost with Economies and Diseconomies of Scale

The long-run average cost curve LAC is the envelope of the short-run average cost curves SAC_1, SAC_2, and SAC_3. With economies and diseconomies of scale, the minimum points of the short-run average cost curves do not lie on the long-run average cost curve.

an output of q_1, its short run average cost would still be \$8.) However, if it expects to produce q_2, the middle-size plant is best. Similarly, with an output of q_3, the largest of the three plants would be the most efficient choice.

What is the firm's long-run average cost curve? In the long run, the firm can change the size of its plant. In doing so, it will always choose the plant that minimizes the average cost of production.

The long-run average cost curve is given by the crosshatched portions of the short-run average cost curves because these show the minimum cost of production for any output level. The long-run average cost curve is the *envelope* of the short-run average cost curves—it envelops or surrounds the short-run curves.

Now suppose that there are many choices of plant size, each having a different short-run average cost curve. Again, the long-run average cost curve is the envelope of the short-run curves. In Figure 7.9 it is the curve LAC. Whatever the firm wants to produce, it can choose the plant size (and the mix of capital and labor) that allows it to produce that output at the minimum average cost. The long-run average cost curve exhibits economies of scale initially but exhibits diseconomies at higher output levels.

To clarify the relationship between short-run and long-run cost curves, consider a firm that wants to produce output q_1. If it builds a small plant, the short-run average cost curve SAC_1 is relevant. The average cost of production (at B on SAC_1) is \$8. A small plant is a better choice than a medium-sized plant with an average cost of production of \$10 (A on curve SAC_2). Point B would therefore become one point on the long-run cost function when only three plant sizes are possible. If plants of other sizes could be built, and if at least one size allowed the firm to produce q_1 at less than \$8 per unit, then B would no longer be on the long-run cost curve.

In Figure 7.9, the envelope that would arise if plants of any size could be built is U-shaped. Note, once again, that the LAC curve never lies above any of the short-run average cost curves. Also note that because there are economies and diseconomies of scale in the long run, the points of minimum average cost of the smallest and largest plants do *not* lie on the long-run average cost curve. For example, a small plant operating at minimum average cost is not efficient because a larger plant can take advantage of increasing returns to scale to produce at a lower average cost.

Finally, note that the long-run marginal cost curve LMC is not the envelope of the short-run marginal cost curves. Short-run marginal costs apply to a particular plant; long-run marginal costs apply to all possible plant sizes. Each point on the long-run marginal cost curve is the short-run marginal cost associated with the most cost-efficient plant. Consistent with this relationship, SMC_1 intersects LMC in Figure 7.9 at the output level q_0 at which SAC_1 is tangent to LAC.

7.5 PRODUCTION WITH TWO OUTPUTS—ECONOMIES OF SCOPE

Many firms produce more than one product. Sometimes a firm's products are closely linked to one another: A chicken farm, for instance, produces poultry and eggs, an automobile company produces automobiles and trucks, and a university produces teaching and research. At other times, firms produce physically unrelated products. In both cases, however, a firm is likely to enjoy production or cost advantages when it produces two or more products.

These advantages could result from the joint use of inputs or production facilities, joint marketing programs, or possibly the cost savings of a common administration. In some cases, the production of one product yields an automatic and unavoidable by-product that is valuable to the firm. For example, sheet metal manufacturers produce scrap metal and shavings that they can sell.

Product Transformation Curves

To study the economic advantages of joint production, let's consider an automobile company that produces two products, cars and tractors. Both products use capital (factories and machinery) and labor as inputs. Cars and tractors are not typically produced at the same plant, but they do share management resources, and both rely on similar machinery and skilled labor. The managers of the company must choose how much of each product to produce. Figure 7.10 shows two **product transformation curves**, each showing the various combinations of cars and tractors that can be produced with a given input of labor and machinery. Curve O_1 describes all combinations of the two outputs that can be produced with a relatively low level of inputs, and curve O_2 describes the output combinations associated with twice the inputs.

• **product transformation curve** Curve showing the various combinations of two different outputs (products) that can be produced with a given set of inputs.

Why does the product transformation curve have a negative slope? Because in order to get more of one output, the firm must give up some of the other output. For example, a firm that emphasizes car production will devote less of its resources to producing tractors. In Figure 7.10, curve O_2 lies twice as far from the origin as curve O_1, signifying that this firm's production process exhibits constant returns to scale in the production of both commodities.

If curve O_1 were a straight line, joint production would entail no gains (or losses). One smaller company specializing in cars and another in tractors

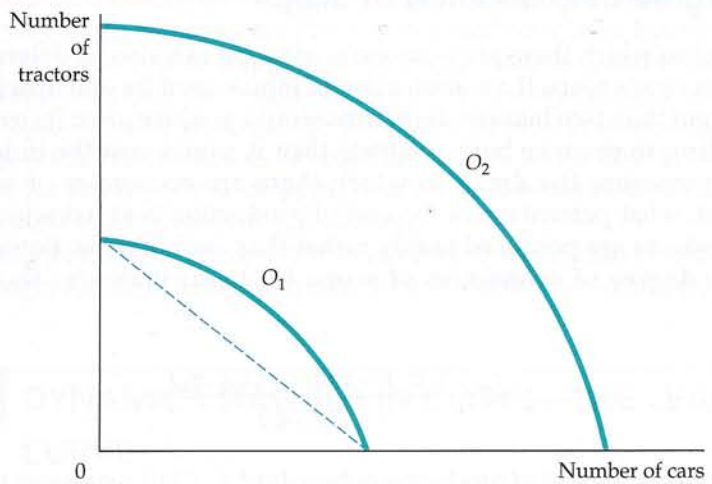

FIGURE 7.10 Product Transformation Curve

The product transformation curve describes the different combinations of two outputs that can be produced with a fixed amount of production inputs. The product transformation curves O_1 and O_2 are bowed out (or concave) because there are economies of scope in production.

would generate the same output as a single company producing both. However, the product transformation curve is bowed outward (or *concave*) because joint production usually has advantages that enable a single company to produce more cars and tractors with the same resources than would two companies producing each product separately. These production advantages involve the joint sharing of inputs. A single management, for example, is often able to schedule and organize production and to handle accounting and financial activities more effectively than separate managements.

Economies and Diseconomies of Scope

• **economies of scope** Situation in which joint output of a single firm is greater than output that could be achieved by two different firms when each produces a single product.

• **diseconomies of scope** Situation in which joint output of a single firm is less than could be achieved by separate firms when each produces a single product.

In general, **economies of scope** are present when the joint output of a single firm is greater than the output that could be achieved by two different firms each producing a single product (with equivalent production inputs allocated between them). If a firm's joint output is *less* than that which could be achieved by separate firms, then its production process involves **diseconomies of scope**. This possibility could occur if the production of one product somehow conflicted with the production of the second.

There is no direct relationship between economies of scale and economies of scope. A two-output firm can enjoy economies of scope even if its production process involves diseconomies of scale. Suppose, for example, that manufacturing flutes and piccolos jointly is cheaper than producing both separately. Yet the production process involves highly skilled labor and is most effective if undertaken on a small scale. Likewise, a joint-product firm can have economies of scale for each individual product yet not enjoy economies of scope. Imagine, for example, a large conglomerate that owns several firms that produce efficiently on a large scale but that do not take advantage of economies of scope because they are administered separately.

The Degree of Economies of Scope

The extent to which there are economies of scope can also be determined by studying a firm's costs. If a combination of inputs used by one firm generates more output than two independent firms would produce, then it costs less for a single firm to produce both products than it would cost the independent firms. To measure the *degree* to which there are economies of scope, we should ask what percentage of the cost of production is saved when two (or more) products are produced jointly rather than individually. Equation (7.7) gives the **degree of economies of scope (SC)** that measures this savings in cost:

• **degree of economies of scope (SC)** Percentage of cost savings resulting when two or more products are produced jointly rather than individually.

$$SC = \frac{C(q_1) + C(q_2) - C(q_1, q_2)}{C(q_1, q_2)} \tag{7.7}$$

$C(q_1)$ represents the cost of producing only output q_1, $C(q_2)$ represents the cost of producing only output q_2, and $C(q_1, q_2)$ the joint cost of producing both outputs. When the physical units of output can be added, as in the car–tractor example, the expression becomes $C(q_1 + q_2)$. With economies of scope, the joint cost is less than the sum of the individual costs. Thus, SC is greater than 0. With diseconomies of scope, SC is negative. In general, the larger the value of SC, the greater the economies of scope.

EXAMPLE 7.5 Economies of Scope in the Trucking Industry

Suppose that you are managing a trucking firm that hauls loads of different sizes between cities.[9] In the trucking business, several related but distinct products can be offered, depending on the size of the load and the length of the haul. First, any load, small or large, can be taken directly from one location to another without intermediate stops. Second, a load can be combined with other loads (which may go between different locations) and eventually be shipped indirectly from its origin to the appropriate destination. Each type of load, partial or full, may involve different lengths of haul.

This range of possibilities raises questions about both economies of scale and economies of scope. The scale question asks whether large-scale, direct hauls are cheaper and more profitable than individual hauls by small truckers. The scope question asks whether a large trucking firm enjoys cost advantages in operating both direct quick hauls and indirect, slower (but less expensive) hauls. Central planning and organization of routes could provide for economies of scope. The key to the presence of economies of scale is the fact that the organization of routes and the types of hauls we have described can be accomplished more efficiently when many hauls are involved. In such cases, a firm is more likely to be able to schedule hauls in which most truckloads are full rather than half-full.

Studies of the trucking industry show that economies of scope are present. For example, one analysis of 105 trucking firms looked at four distinct outputs: (1) short hauls with partial loads, (2) intermediate hauls with partial loads, (3) long hauls with partial loads, and (4) hauls with total loads. The results indicate that the degree of economies of scope SC was 1.576 for a reasonably large firm. However, the degree of economies of scope falls to 0.104 when the firm becomes very large. Because large firms carry sufficiently large truckloads, there is usually no advantage to stopping at an intermediate terminal to fill a partial load. A direct trip from the origin to the destination is sufficient. Apparently, however, because other disadvantages are associated with the management of very large firms, the economies of scope get smaller as the firm gets bigger. In any event, the ability to combine partial loads at an intermediate location lowers the firm's costs and increases its profitability.

The study suggests, therefore, that to compete in the trucking industry, a firm must be large enough to be able to combine loads at intermediate stopping points.

*7.6 DYNAMIC CHANGES IN COSTS—THE LEARNING CURVE

Our discussion thus far has suggested one reason why a large firm may have a lower long-run average cost than a small firm: increasing returns to scale in production. It is tempting to conclude that firms that enjoy lower average cost over

[9]This example is based on Judy S. Wang Chiang and Ann F. Friedlaender, "Truck Technology and Efficient Market Structure," *Review of Economics and Statistics* 67 (1985): 250–58.

time are growing firms with increasing returns to scale. But this need not be true. In some firms, long-run average cost may decline over time because workers and managers absorb new technological information as they become more experienced at their jobs.

As management and labor gain experience with production, the firm's marginal and average costs of producing a given level of output fall for four reasons:

1. Workers often take longer to accomplish a given task the first few times they do it. As they become more adept, their speed increases.
2. Managers learn to schedule the production process more effectively, from the flow of materials to the organization of the manufacturing itself.
3. Engineers who are initially cautious in their product designs may gain enough experience to be able to allow for tolerances in design that save costs without increasing defects. Better and more specialized tools and plant organization may also lower cost.
4. Suppliers may learn how to process required materials more effectively and pass on some of this advantage in the form of lower costs.

As a consequence, a firm "learns" over time as cumulative output increases. Managers can use this learning process to help plan production and forecast future costs. Figure 7.11 illustrates this process in the form of a **learning curve**—a curve that describes the relationship between a firm's cumulative output and the amount of inputs needed to produce each unit of output.

• **learning curve** Graph relating amount of inputs needed by a firm to produce each unit of output to its cumulative output.

Graphing the Learning Curve

Figure 7.11 shows a learning curve for the production of machine tools. The horizontal axis measures the *cumulative* number of lots of machine tools (groups of approximately 40) that the firm has produced. The vertical axis shows the

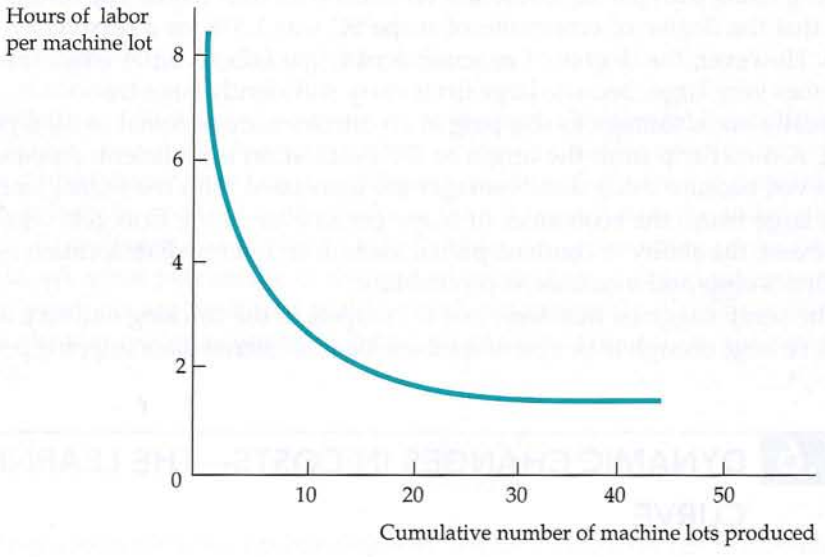

FIGURE 7.11 The Learning Curve

A firm's production cost may fall over time as managers and workers become more experienced and more effective at using the available plant and equipment. The learning curve shows the extent to which hours of labor needed per unit of output fall as the cumulative output increases.

number of hours of labor needed to produce each lot. Labor input per unit of output directly affects the production cost because the fewer the hours of labor needed, the lower the marginal and average cost of production.

The learning curve in Figure 7.11 is based on the relationship

$$L = A + BN^{-\beta} \tag{7.8}$$

where N is the cumulative units of output produced and L the labor input per unit of output. A, B, and β are constants, with A and B positive, and β between 0 and 1. When N is equal to 1, L is equal to $A + B$, so that $A + B$ measures the labor input required to produce the first unit of output. When β equals 0, labor input per unit of output remains the same as the cumulative level of output increases; there is no learning. When β is positive and N gets larger and larger, L becomes arbitrarily close to A. A, therefore, represents the minimum labor input per unit of output after all learning has taken place.

The larger β is, the more important the learning effect. With β equal to 0.5, for example, the labor input per unit of output falls proportionately to the square root of the cumulative output. This degree of learning can substantially reduce production costs as a firm becomes more experienced.

In this machine tool example, the value of β is 0.31. For this particular learning curve, every doubling in cumulative output causes the input requirement (less the minimum attainable input requirement) to fall by about 20 percent.[10] As Figure 7.11 shows, the learning curve drops sharply as the cumulative number of lots increases to about 20. Beyond an output of 20 lots, the cost savings are relatively small.

Learning versus Economies of Scale

Once the firm has produced 20 or more machine lots, the entire effect of the learning curve would be complete, and we could use the usual analysis of cost. If, however, the production process were relatively new, relatively high cost at low levels of output (and relatively low cost at higher levels) would indicate learning effects, not economies of scale. With learning, the cost of production for a mature firm is relatively low regardless of the scale of the firm's operation. If a firm that produces machine tools in lots knows that it enjoys economies of scale, it should produce its machines in very large lots to take advantage of the lower cost associated with size. If there is a learning curve, the firm can lower its cost by scheduling the production of many lots regardless of individual lot size.

Figure 7.12 shows this phenomenon. AC_1 represents the long-run average cost of production of a firm that enjoys economies of scale in production. Thus the increase in the rate of output from A to B along AC_1 leads to lower cost due to economies of scale. However, the move from A on AC_1 to C on AC_2 leads to lower cost due to learning, which shifts the average cost curve downward.

The learning curve is crucial for a firm that wants to predict the cost of producing a new product. Suppose, for example, that a firm producing machine tools knows that its labor requirement per machine for the first 10 machines is 1.0, the minimum labor requirement A is equal to zero, and β is approximately equal to 0.32. Table 7.3 calculates the total labor requirement for producing 80 machines.

Because there is a learning curve, the per-unit labor requirement falls with increased production. As a result, the total labor requirement for producing more and more output increases in smaller and smaller increments.

[10] Because $(L - A) = BN^{-.31}$, we can check that $0.8(L - A)$ is approximately equal to $B(2N)^{-.31}$.

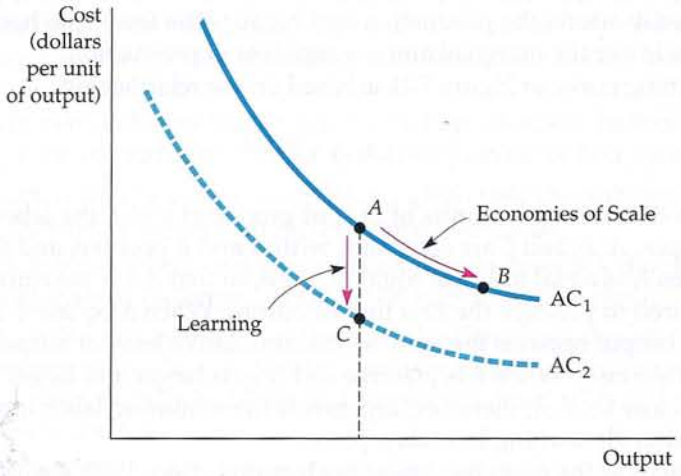

FIGURE 7.12 Economies of Scale versus Learning

A firm's average cost of production can decline over time because of growth of sales when increasing returns are present (a move from *A* to *B* on curve AC_1), or it can decline because there is a learning curve (a move from *A* on curve AC_1 to *C* on curve AC_2).

Therefore, a firm looking only at the high initial labor requirement will obtain an overly pessimistic view of the business. Suppose the firm plans to be in business for a long time, producing 10 units per year. Suppose the total labor requirement for the first year's production is 10. In the first year of production, the firm's cost will be high as it learns the business. But once the learning effect has taken place, production costs will fall. After 8 years, the labor required to produce 10 units will be only 5.1, and per-unit cost will be roughly half what it was in the first year of production. Thus, the learning curve can be important for a firm deciding whether it is profitable to enter an industry.

TABLE 7.3 Predicting the Labor Requirements of Producing a Given Output

Cumulative Output (*N*)	Per-Unit Labor Requirement for Each 10 Units of Output (*L*)*	Total Labor Requirement
10	1.00	10.0
20	.80	18.0 = (10.0 + 8.0)
30	.70	25.0 = (18.0 + 7.0)
40	.64	31.4 = (25.0 + 6.4)
50	.60	37.4 = (31.4 + 6.0)
60	.56	43.0 = (37.4 + 5.6)
70	.53	48.3 = (43.0 + 5.3)
80	.51	53.4 = (48.3 + 5.1)

*The numbers in this column were calculated from the equation $\log(L) = -0.322 \log(N/10)$, where *L* is the unit labor input and *N* is cumulative output.

EXAMPLE 7.6 The Learning Curve in Practice

Suppose that as the manager of a firm that has just entered the chemical processing industry, you face the following problem: Should you produce a relatively low level of output and sell at a high price, or should you price your product lower and increase your rate of sales? The second alternative is appealing if there is a learning curve in the industry: The increased volume will lower your average production costs over time and increase profitability.

To decide what to do, you can examine the available statistical evidence that distinguishes the components of the learning curve (learning new processes by labor, engineering improvements, etc.) from increasing returns to scale. For example, a study of 37 chemical products reveals that cost reductions in the chemical processing industry are directly tied to the growth of cumulative industry output, to investment in improved capital equipment, and, to a lesser extent, to economies of scale.[11] In fact, for the entire sample of chemical products, average costs of production fall at 5.5 percent per year. The study reveals that for each doubling of plant scale, the average cost of production falls by 11 percent. For each doubling of cumulative output, however, the average cost of production falls by 27 percent. The evidence shows clearly that learning effects are more important than economies of scale in the chemical processing industry.[12]

The learning curve has also been shown to be important in the semiconductor industry. A study of seven generations of dynamic random-access memory (DRAM) semiconductors from 1974 to 1992 found that the learning rates averaged about 20 percent; thus a 10-percent increase in cumulative production would lead to a 2-percent decrease in cost.[13] The study also compared learning by firms in Japan to firms in the United States and found that there was no distinguishable difference in the speed of learning.

Another example is the aircraft industry, where studies have found learning rates that are as high as 40 percent. This is illustrated in Figure 7.13, which shows the labor requirements for producing aircraft by Airbus Industrie. Observe that the first 10 or 20 airplanes require far more labor to produce than the hundredth or two hundredth airplane. Also note how the learning curve flattens out after a certain point; in this case nearly all learning is complete after 200 airplanes have been built.

---

[11] The study was conducted by Marvin Lieberman, "The Learning Curve and Pricing in the Chemical Processing Industries," *RAND Journal of Economics* 15 (1984): 213–28.

[12] The author used the average cost AC of the chemical products, the cumulative industry output X, and the average scale of a production plant Z. He then estimated the relationship $\log(AC) = -0.387 \log(X) - 0.173 \log(Z)$. The -0.387 coefficient on cumulative output tells us that for every 1-percent increase in cumulative output, average cost decreases 0.387 percent. The -0.173 coefficient on plant size tells us that for every 1-percent increase in plant size, average cost decreases 0.173 percent.

By interpreting the two coefficients in light of the output and plant-size variables, we can allocate about 15 percent of the cost reduction to increases in the average scale of plants and 85 percent to increases in cumulative industry output. Suppose plant scale doubled while cumulative output increased by a factor of 5 during the study. In that case, costs would fall by 11 percent from the increased scale and by 62 percent from the increase in cumulative output.

[13] The study was conducted by D. A. Irwin and P. J. Klenow, "Learning-by-Doing Spillovers in the Semiconductor Industry," *Journal of Political Economy* 102 (December 1994): 1200–27.

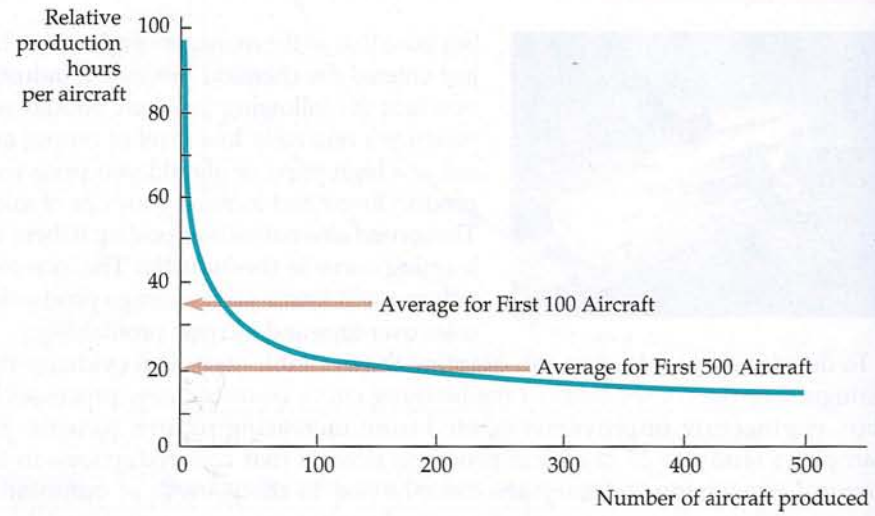

FIGURE 7.13 Learning Curve for Airbus Industrie

The learning curve relates the labor requirement per aircraft to the cumulative number of aircraft produced. As the production process becomes better organized and workers gain familiarity with their jobs, labor requirements fall dramatically.

Learning-curve effects can be important in determining the shape of long-run cost curves and can thus help guide management decisions. Managers can use learning-curve information to decide whether a production operation is profitable and, if so, how to plan how large the plant operation and the volume of cumulative output need be to generate a positive cash flow.

*7.7 ESTIMATING AND PREDICTING COST

• **cost function** Function relating cost of production to level of output and other variables that the firm can control.

A business that is expanding or contracting its operation must predict how costs will change as output changes. Estimates of future costs can be obtained from a **cost function**, which relates the cost of production to the level of output and other variables that the firm can control.

Suppose we wanted to characterize the short-run cost of production in the automobile industry. We could obtain data on the number of automobiles Q produced by each car company and relate this information to the company's variable cost of production VC. The use of variable cost, rather than total cost, avoids the problem of trying to allocate the fixed cost of a multiproduct firm's production process to the particular product being studied.[14]

Figure 7.14 shows a typical pattern of cost and output data. Each point on the graph relates the output of an auto company to that company's variable cost of production. To predict cost accurately, we must determine the underlying relationship between variable cost and output. Then, if a company expands its

[14]If an additional piece of equipment is needed as output increases, then the annual rental cost of the equipment should be counted as a variable cost. If, however, the same machine can be used at all output levels, its cost is fixed and should not be included.

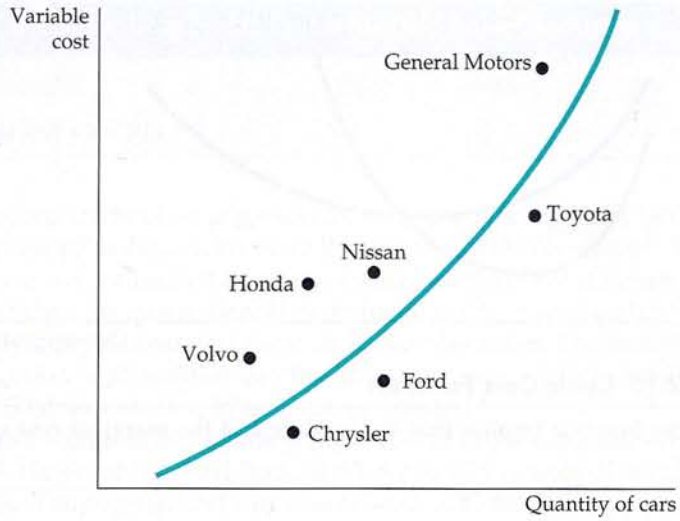

FIGURE 7.14 Variable Cost Curve for the Automobile Industry

An empirical estimate of the variable cost curve can be obtained by using data for individual firms in an industry. The variable cost curve for automobile production is obtained by determining statistically the curve that best fits the points that relate the output of each firm to the firm's variable cost of production.

production, we can calculate what the associated cost is likely to be. The curve in the figure is drawn with this in mind—it provides a reasonably close fit to the cost data. (Typically, least-squares regression analysis would be used to fit the curve to the data.) But what shape is the most appropriate, and how do we represent that shape algebraically?

Least-squares regression is explained in the appendix to this book.

Here is one cost function that we might choose:

$$VC = \beta q \tag{7.9}$$

Although easy to use, this *linear* relationship between cost and output is applicable only if marginal cost is constant.[15] For every unit increase in output, variable cost increases by β; marginal cost is thus constant and equal to β.

If we wish to allow for a U-shaped average cost curve and a marginal cost that is not constant, we must use a more complex cost function. One possibility is the *quadratic* cost function, which relates variable cost to output and output squared:

$$VC = \beta q + \gamma q^2 \tag{7.10}$$

This function implies a straight-line marginal cost curve of the form $MC = \beta + 2\gamma q$.[16] Marginal cost increases with output if γ is positive and decreases with output if γ is negative.

If the marginal cost curve is not linear, we might use a *cubic* cost function:

$$VC = \beta q + \gamma q^2 + \delta q^3 \tag{7.11}$$

[15]In statistical cost analyses, other variables might be added to the cost function to account for differences in input costs, production processes, production mix, etc., among firms.

[16]Short-run marginal cost is given by $\Delta VC/\Delta q = \beta + \gamma\Delta(q^2)$. But $\Delta(q^2)/\Delta q = 2q$. (Check this by using calculus or by numerical example.) Therefore, $MC = \beta + 2\gamma q$.

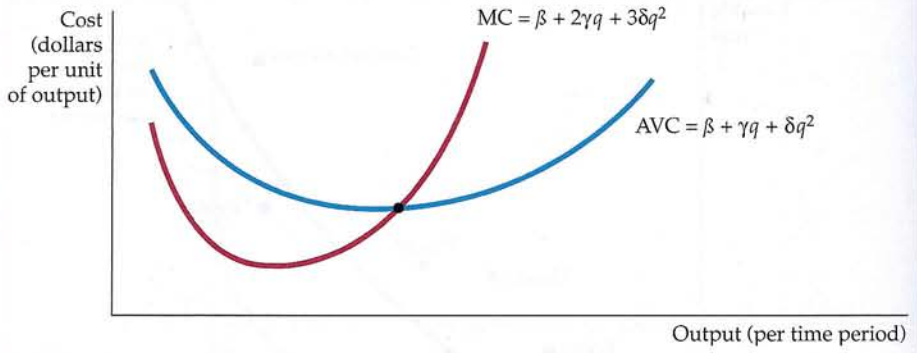

FIGURE 7.15 Cubic Cost Function

A cubic cost function implies that the average and the marginal cost curves are U-shaped.

Figure 7.15 shows this cubic cost function. It implies U-shaped marginal as well as average cost curves.

Cost functions can be difficult to measure for several reasons. First, output data often represent an aggregate of different types of products. The automobiles produced by General Motors, for example, involve different models of cars. Second, cost data are often obtained directly from accounting information that fails to reflect opportunity costs. Third, allocating maintenance and other plant costs to a particular product is difficult when the firm is a conglomerate that produces more than one product line.

Cost Functions and the Measurement of Scale Economies

Recall that the cost-output elasticity E_C is less than one when there are economies of scale and greater than one when there are diseconomies of scale. The *scale economies index (SCI)* provides an index of whether or not there are scale economies. SCI is defined as follows:

$$\text{SCI} = 1 - E_C \tag{7.12}$$

When $E_C = 1$, SCI = 0 and there are no economies or diseconomies of scale. When E_C is greater than one, SCI is negative and there are diseconomies of scale. Finally, when E_C is less than 1, SCI is positive and there are economies of scale.

EXAMPLE 7.7 Cost Functions for Electric Power

In 1955, consumers bought 369 billion kilowatt-hours (kwh) of electricity; in 1970 they bought 1083 billion. Because there were fewer electric utilities in 1970, the output per firm had increased substantially. Was this increase due to economies of scale or to other factors? If it was the result of economies of scale, it would be economically inefficient for regulators to "break up" electric utility monopolies.

TABLE 7.4 Scale Economies in the Electric Power Industry

Output (million kwh)	43	338	1109	2226	5819
Value of SCI, 1955	.41	.26	.16	.10	.04

An interesting study of scale economies was based on the years 1955 and 1970 for investor-owned utilities with more than $1 million in revenues.[17] The cost of electric power was estimated by using a cost function that is somewhat more sophisticated than the quadratic and cubic functions discussed earlier.[18] Table 7.4 shows the resulting estimates of the scale economies index. The results are based on a classification of all utilities into five size categories, with the median output (measured in kilowatt-hours) in each category listed.

The positive values of SCI tell us that all sizes of firms had some economies of scale in 1955. However, the magnitude of the economies of scale diminishes as firm size increases. The average cost curve associated with the 1955 study is drawn in Figure 7.16 and labeled 1955. The point of minimum average cost occurs at point *A*, at an output of approximately 20 billion kilowatts. Because there were no firms of this size in 1955, no firm had exhausted the opportunity for returns to scale in production. Note, however, that the average cost curve is relatively flat from an output of 9 billion kilowatts and higher, a range in which 7 of 124 firms produced.

When the same cost functions were estimated with 1970 data, the cost curve labeled 1970 in Figure 7.16 was the result. The graph shows clearly that the

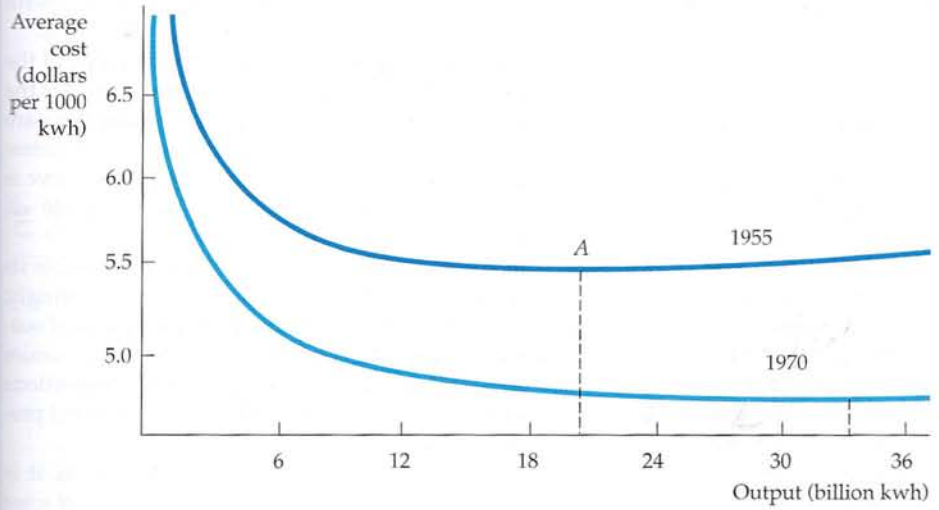

FIGURE 7.16 Average Cost of Production in the Electric Power Industry

The average cost of electric power in 1955 achieved a minimum at approximately 20 billion kilowatt-hours. By 1970 the average cost of production had fallen sharply and achieved a minimum at an output of more than 33 billion kilowatt-hours.

[17] This example is based on Laurits Christensen and William H. Greene, "Economies of Scale in U.S. Electric Power Generation," *Journal of Political Economy* 84 (1976): 655–76.

[18] The translog cost function used in this study provides a more general functional relationship than any of those we have discussed.

average costs of production fell from 1955 to 1970. (The data are in real 1970 dollars.) But the flat part of the curve now begins at about 15 billion kwh. By 1970, 24 of 80 firms were producing in this range. Thus, many more firms were operating in the flat portion of the average cost curve in which economies of scale are not an important phenomenon. More important, most of the firms were producing in a portion of the 1970 cost curve that was flatter than their point of operation on the 1955 curve. (Five firms were at points of diseconomies of scale: Consolidated Edison [SCI = –0.003], Detroit Edison [SCI = –0.004], Duke Power [SCI = –0.012], Commonwealth Edison [SCI = –0.014], and Southern [SCI = –0.028].) Thus, unexploited scale economies were much smaller in 1970 than in 1955.

This cost function analysis makes it clear that the decline in the cost of producing electric power cannot be explained by the ability of larger firms to take advantage of economies of scale. Rather, improvements in technology unrelated to the scale of the firms' operation and the decline in the real cost of energy inputs, such as coal and oil, are important reasons for the lower costs. The tendency toward lower average cost reflecting a movement to the right along an average cost curve is minimal compared with the effect of technological improvement.

SUMMARY

1. Managers, investors, and economists must take into account the *opportunity cost* associated with the use of a firm's resources: the cost associated with the opportunities forgone when the firm uses its resources in its next best alternative.
2. A *sunk cost* is an expenditure that has been made and cannot be recovered. After it has been incurred, it should be ignored when making future economic decisions.
3. In the short run, one or more of a firm's inputs are fixed. Total cost can be divided into fixed cost and variable cost. A firm's *marginal cost* is the additional variable cost associated with each additional unit of output. The *average variable cost* is the total variable cost divided by the number of units of output.
4. In the short run, when not all inputs are variable, the presence of diminishing returns determines the shape of the cost curves. In particular, there is an inverse relationship between the marginal product of a single variable input and the marginal cost of production. The average variable cost and average total cost curves are U-shaped. The short-run marginal cost curve increases beyond a certain point, and cuts both average cost curves from below at their minimum points.
5. In the long run, all inputs to the production process are variable. As a result, the choice of inputs depends both on the relative costs of the factors of production and on the extent to which the firm can substitute among inputs in its production process. The cost-minimizing input choice is made by finding the point of tangency between the isoquant representing the level of desired output and an isocost line.
6. The firm's *expansion path* shows how its cost-minimizing input choices vary as the scale or output of its operation increases. As a result, the expansion path provides useful information relevant for long-run planning decisions.
7. The long-run average cost curve is the envelope of the firm's short-run average cost curves, and it reflects the presence or absence of returns to scale. When there are increasing returns to scale initially and then decreasing returns to scale, the long-run average cost curve is U-shaped, and the envelope does not include all points of minimum short-run average cost.
8. A firm enjoys *economies of scale* when it can double its output at less than twice the cost. Correspondingly, there are diseconomies of scale when a doubling of output requires more than twice the cost. Scale economies and diseconomies apply even when input proportions are variable; returns to scale apply only when input proportions are fixed.
9. When a firm produces two (or more) outputs, it is important to note whether there are *economies of scope* in production. Economies of scope arise when the firm can produce any combination of the two outputs more cheaply than could two independent firms that each produced a single output. The degree of economies of scope is measured by the percentage reduction in cost when one firm produces two products relative to the cost of producing them individually.
10. A firm's average cost of production can fall over time if the firm "learns" how to produce more effectively. The *learning curve* shows how much the input needed to produce a given output falls as the cumulative output of the firm increases.

11. Cost functions relate the cost of production to the firm's level of output. The functions can be measured in both the short run and the long run by using either data for firms in an industry at a given time or data for an industry over time. A number of functional relationships, including linear, quadratic, and cubic, can be used to represent cost functions.

QUESTIONS FOR REVIEW

1. A firm pays its accountant an annual retainer of $10,000. Is this an economic cost?
2. The owner of a small retail store does her own accounting work. How would you measure the opportunity cost of her work?
3. Please explain whether the following statements are true or false.
   a. If the owner of a business pays himself no salary, then the accounting cost is zero, but the economic cost is positive.
   b. A firm that has positive accounting profit does not necessarily have positive economic profit.
   c. If a firm hires a currently unemployed worker, the opportunity cost of utilizing the worker's services is zero.
4. Suppose that labor is the only variable input to the production process. If the marginal cost of production is diminishing as more units of output are produced, what can you say about the marginal product of labor?
5. Suppose a chair manufacturer finds that the marginal rate of technical substitution of capital for labor in her production process is substantially greater than the ratio of the rental rate on machinery to the wage rate for assembly-line labor. How should she alter her use of capital and labor to minimize the cost of production?
6. Why are isocost lines straight lines?
7. Assume that the marginal cost of production is increasing. Can you determine whether the average variable cost is increasing or decreasing? Explain.
8. Assume that the marginal cost of production is greater than the average variable cost. Can you determine whether the average variable cost is increasing or decreasing? Explain.
9. If the firm's average cost curves are U-shaped, why does its average variable cost curve achieve its minimum at a lower level of output than the average total cost curve?
10. If a firm enjoys economies of scale up to a certain output level, and cost then increases proportionately with output, what can you say about the shape of the long-run average cost curve?
11. How does a change in the price of one input change the firm's long-run expansion path?
12. Distinguish between economies of scale and economies of scope. Why can one be present without the other?
13. Is the firm's expansion path always a straight line?
14. What is the difference between economies of scale and returns to scale?

EXERCISES

1. Joe quits his computer programming job, where he was earning a salary of $50,000 per year, to start his own computer software business in a building that he owns and was previously renting out for $24,000 per year. In his first year of business he has the following expenses: salary paid to himself, $40,000; rent, $0; other expenses, $25,000. Find the accounting cost and the economic cost associated with Joe's computer software business.
2. a. Fill in the blanks in the table on page 262.
   b. Draw a graph that shows marginal cost, average variable cost, and average total cost, with cost on the vertical axis and quantity on the horizontal axis.
3. A firm has a fixed production cost of $5000 and a constant marginal cost of production of $500 per unit produced.
   a. What is the firm's total cost function? Average cost?
   b. If the firm wanted to minimize the average total cost, would it choose to be very large or very small? Explain.
4. Suppose a firm must pay an annual tax, which is a fixed sum, independent of whether it produces any output.
   a. How does this tax affect the firm's fixed, marginal, and average costs?
   b. Now suppose the firm is charged a tax that is proportional to the number of items it produces. Again, how does this tax affect the firm's fixed, marginal, and average costs?
5. A recent issue of *Business Week* reported the following:

   > During the recent auto sales slump, GM, Ford, and Chrysler decided it was cheaper to sell cars to rental companies at a loss than to lay off workers. That's because closing and reopening plants is expensive, partly because the auto makers' current union contracts obligate them to pay many workers even if they're not working.

   When the article discusses selling cars "at a loss," is it referring to accounting profit or economic profit? How will the two differ in this case? Explain briefly.

Units of Output	Fixed Cost	Variable Cost	Total Cost	Marginal Cost	Average Fixed Cost	Average Variable Cost	Average Total Cost
0			100				
1			125				
2			145				
3			157				
4			177				
5			202				
6			236				
7			270				
8			326				
9			398				
10			490				

6. Suppose the economy takes a downturn, and that labor costs fall by 50 percent and are expected to stay at that level for a long time. Show graphically how this change in the relative price of labor and capital affects the firm's expansion path.
7. The cost of flying a passenger plane from point A to point B is \$50,000. The airline flies this route four times per day at 7 AM, 10 AM, 1 PM, and 4 PM. The first and last flights are filled to capacity with 240 people. The second and third flights are only half full. Find the average cost per passenger for each flight. Suppose the airline hires you as a marketing consultant and wants to know which type of customer it should try to attract—the off-peak customer (the middle two flights) or the rush-hour customer (the first and last flights). What advice would you offer?
8. You manage a plant that mass-produces engines by teams of workers using assembly machines. The technology is summarized by the production function

$$q = 5KL$$

where q is the number of engines per week, K is the number of assembly machines, and L is the number of labor teams. Each assembly machine rents for r = \$10,000 per week, and each team costs w = \$5000 per week. Engine costs are given by the cost of labor teams and machines, plus \$2000 per engine for raw materials. Your plant has a fixed installation of 5 assembly machines as part of its design.
   a. What is the cost function for your plant—namely, how much would it cost to produce q engines? What are average and marginal costs for producing q engines? How do average costs vary with output?
   b. How many teams are required to produce 250 engines? What is the average cost per engine?
   c. You are asked to make recommendations for the design of a new production facility. What capital/labor (K/L) ratio should the new plant accommodate if it wants to minimize the total cost of producing at any level of output q?
9. The short-run cost function of a company is given by the equation TC = 200 + 55q, where TC is the total cost and q is the total quantity of output, both measured in thousands.
   a. What is the company's fixed cost?
   b. If the company produced 100,000 units of goods, what would be its average variable cost?
   c. What would be its marginal cost of production?
   d. What would be its average fixed cost?
   e. Suppose the company borrows money and expands its factory. Its fixed cost rises by \$50,000, but its variable cost falls to \$45,000 per 1000 units. The cost of interest (i) also enters into the equation. Each 1-point increase in the interest rate raises costs by \$3000. Write the new cost equation.

*10. A chair manufacturer hires its assembly-line labor for \$30 an hour and calculates that the rental cost of its machinery is \$15 per hour. Suppose that a chair can be produced using 4 hours of labor or machinery in any combination. If the firm is currently using 3 hours of labor for each hour of machine time, is it minimizing its costs of production? If so, why? If not, how can it improve the situation? Graphically illustrate the isoquant and the two isocost lines for the current combination of labor and capital and for the optimal combination of labor and capital.

*11. Suppose that a firm's production function is $q = 10L^{\frac{1}{2}}K^{\frac{1}{2}}$. The cost of a unit of labor is \$20 and the cost of a unit of capital is \$80.
   a. The firm is currently producing 100 units of output and has determined that the cost-minimizing

quantities of labor and capital are 20 and 5, respectively. Graphically illustrate this using isoquants and isocost lines.

b. The firm now wants to increase output to 140 units. If capital is fixed in the short run, how much labor will the firm require? Illustrate this graphically and find the firm's new total cost.

c. Graphically identify the cost-minimizing level of capital and labor in the long run if the firm wants to produce 140 units.

d. If the marginal rate of technical substitution is K/L, find the optimal level of capital and labor required to produce the 140 units of output.

***12.** A computer company's cost function, which relates its average cost of production AC to its cumulative output in thousands of computers Q and its plant size in terms of thousands of computers produced per year q (within the production range of 10,000 to 50,000 computers), is given by

$$AC = 10 - 0.1Q + 0.3q$$

a. Is there a learning-curve effect?

b. Are there economies or diseconomies of scale?

c. During its existence, the firm has produced a total of 40,000 computers and is producing 10,000 computers this year. Next year it plans to increase production to 12,000 computers. Will its average cost of production increase or decrease? Explain.

***13.** Suppose the long-run total cost function for an industry is given by the cubic equation $TC = a + bq + cq^2 + dq^3$. Show (using calculus) that this total cost function is consistent with a U-shaped average cost curve for at least some values of a, b, c, and d.

***14.** A computer company produces hardware and software using the same plant and labor. The total cost of producing computer processing units H and software programs S is given by

$$TC = aH + bS - cHS$$

where a, b, and c are positive. Is this total cost function consistent with the presence of economies or diseconomies of scale? With economies or diseconomies of scope?

Appendix to Chapter 7

PRODUCTION AND COST THEORY—A MATHEMATICAL TREATMENT

This appendix presents a mathematical treatment of the basics of production and cost theory. As in the appendix to Chapter 4, we use the method of Lagrange multipliers to solve the firm's cost-minimizing problem.

Cost Minimization

The theory of the firm relies on the assumption that firms choose inputs to the production process that minimize the cost of producing output. If there are two inputs, capital K and labor L, the production function $F(K, L)$ describes the maximum output that can be produced for every possible combination of inputs. We assume that each factor in the production process has positive but decreasing marginal products. Therefore, writing the marginal product of capital and labor as $MP_K(K, L)$ and $MP_L(K, L)$, respectively, it follows that

$$MP_K(K, L) = \frac{\partial F(K, L)}{\partial K} > 0, \quad \frac{\partial^2 F(K, L)}{\partial K^2} < 0$$

$$MP_L(K, L) = \frac{\partial F(K, L)}{\partial L} > 0, \quad \frac{\partial^2 F(K, L)}{\partial L^2} < 0$$

A competitive firm takes the prices of both labor w and capital r as given. Then the cost-minimization problem can be written as

$$\text{Minimize } C = wL + rK \tag{A7.1}$$

subject to the constraint that a fixed output q_0 be produced:

$$F(K, L) = q_0 \tag{A7.2}$$

C represents the cost of producing the fixed level of output q_0.

To determine the firm's demand for capital and labor inputs, we choose the values of K and L that minimize (A7.1) subject to (A7.2). We can solve this constrained optimization problem in three steps using the method discussed in the appendix to Chapter 4:

- ***Step 1:*** Set up the Lagrangian, which is the sum of two components: the cost of production (to be minimized) and the Lagrange multiplier λ times the output constraint faced by the firm:

$$\Phi = wL + rK - \lambda[F(K, L) - q_0] \tag{A7.3}$$

- ***Step 2:*** Differentiate the Lagrangian with respect to K, L, and λ. Then equate the resulting derivatives to zero to obtain the necessary conditions for a minimum.[1]

$$\begin{aligned} \partial\Phi / \partial K &= r - \lambda \mathrm{MP}_K(K, L) = 0 \\ \partial\Phi / \partial L &= w - \lambda \mathrm{MP}_L(K, L) = 0 \\ \partial\Phi / \partial \lambda &= q_0 - F(K, L) = 0 \end{aligned} \tag{A7.4}$$

- ***Step 3:*** In general, these equations can be solved to obtain the optimizing values of L, K, and λ. It is particularly instructive to combine the first two conditions in (A7.4) to obtain

$$\mathrm{MP}_K(K, L)/r = \mathrm{MP}_L(K, L)/w \tag{A7.5}$$

Equation (A7.5) tells us that if the firm is minimizing costs, it will choose its factor inputs to equate the ratio of the marginal product of each factor divided by its price. This is exactly the same condition that we derived as Equation 7.4 (page 238) in the text.

Finally, we can rewrite the first two conditions of (A7.4) to evaluate the Lagrange multiplier:

$$\begin{aligned} r - \lambda \mathrm{MP}_K(K, L) = 0 \Rightarrow \lambda &= \frac{r}{\mathrm{MP}_K(K, L)} \\ w - \lambda \mathrm{MP}_L(K, L) = 0 \Rightarrow \lambda &= \frac{w}{\mathrm{MP}_L(K, L)} \end{aligned} \tag{A7.6}$$

Suppose output increases by one unit. Because the marginal product of capital measures the extra output associated with an additional input of capital, $1/\mathrm{MP}_K(K, L)$ measures the extra capital needed to produce one unit of output. Therefore, $r/\mathrm{MP}_K(K, L)$ measures the additional input cost of producing an additional unit of output by increasing capital. Likewise, $w/\mathrm{MP}_L(K, L)$ measures the additional cost of producing a unit of output using additional labor as an input. In both cases, the Lagrange multiplier is equal to the marginal cost of production because it tells us how much the cost increases if the amount produced is increased by one unit.

Marginal Rate of Technical Substitution

Recall that an *isoquant* is a curve that represents the set of all input combinations that give the firm the same level of output—say, q_0. Thus, the condition that $F(K, L) = q_0$ represents a production isoquant. As input combinations are changed along an isoquant, the change in output, given by the total derivative of $F(K, L)$ equals zero (i.e., $\mathrm{d}q = 0$). Thus

$$\mathrm{MP}_K(K, L)\mathrm{d}K + \mathrm{MP}_L(K, L)\mathrm{d}L = \mathrm{d}q = 0 \tag{A7.7}$$

It follows by rearrangement that

$$-\mathrm{d}K/\mathrm{d}L = \mathrm{MRTS}_{LK} = \mathrm{MP}_L(K, L)/\mathrm{MP}_K(K, L) \tag{A7.8}$$

where MRTS_{LK} is the firm's marginal rate of technical substitution between labor and capital.

[1] These conditions are necessary for a solution involving positive amounts of both inputs.

Now, rewrite the condition given by (A7.5) to get

$$MP_L(K, L)/MP_K(K, L) = w/r \quad \textbf{(A7.9)}$$

Because the left side of (A7.8) represents the negative of the slope of the isoquant, it follows that at the point of tangency of the isoquant and the isocost line, the firm's marginal rate of technical substitution (which trades off inputs while keeping output constant) is equal to the ratio of the input prices (which represents the slope of the firm's isocost line).

We can look at this result another way by rewriting (A7.9) again:

$$MP_L/w = MP_K/r \quad \textbf{(A7.10)}$$

Equation (A7.10) is the same as (A7.5) and tells us that the marginal products of all production inputs must be equal when these marginal products are adjusted by the unit cost of each input.

Duality in Production and Cost Theory

As in consumer theory, the firm's input decision has a dual nature. The optimum choice of K and L can be analyzed not only as the problem of choosing the lowest isocost line tangent to the production isoquant, but also as the problem of choosing the highest production isoquant tangent to a given isocost line. Suppose we wish to spend C_0 on production. The dual problem asks what combination of K and L will let us produce the most output at a cost of C_0. We can see the equivalence of the two approaches by solving the following problem:

$$\text{Maximize } F(K, L) \text{ subject to } wL + rL = C_0$$

We can solve this problem using the Lagrangian method:

- ***Step 1:*** We set up the Lagrangian

$$\Phi = F(K, L) - \mu(wL + rK - C_0) \quad \textbf{(A7.12)}$$

  where μ is the Lagrange multiplier.

- ***Step 2:*** We differentiate the Lagrangian with respect to K, L, and μ and set the resulting equation equal to zero to find the necessary conditions for a maximum:

$$\begin{aligned} \frac{\partial \Phi}{\partial K} &= MP_K(K, L) - \mu r = 0 \\ \frac{\partial \Phi}{\partial K} &= MP_L(K, L) - \mu w = 0 \\ \frac{\partial \Phi}{\partial K} &= wL - rK + C_0 = 0 \end{aligned} \quad \textbf{(A7.13)}$$

- ***Step 3:*** Normally, we can use the equations of A7.13 to solve for K and L. In particular, we combine the first two equations to see that

$$\begin{aligned} \mu &= \frac{MP_K(K, L)}{r} \\ \mu &= \frac{MP_L(K, L)}{w} \\ \Rightarrow \frac{MP_K(K, L)}{r} &= \frac{MP_L(K, L)}{w} \end{aligned} \quad \textbf{(A7.14)}$$

This is the same result as A7.5—that is, the necessary condition for cost minimization.

The Cobb-Douglas Cost and Production Functions

Given a specific production function $F(K, L)$, conditions (A7.13) and (A7.14) can be used to derive the *cost function* $C(q)$. To understand this principle, let's work through the example of a **Cobb-Douglas production function**. This production function is

$$F(K, L) = AK^{\alpha}L^{\beta}$$

where A, α, and β are positive constants.

• **Cobb-Douglas production function** Production function of the form $q = AK^{\alpha}L^{\beta}$, where q is the rate of output, K is the quantity of capital, and L is the quantity of labor, and where A, α, and β are positive constants.

We assume that $\alpha < 1$ and $\beta < 1$, so that the firm has decreasing marginal products of labor and capital.[2] If $\alpha + \beta = 1$, the firm has *constant returns to scale*, because doubling K and L doubles F. If $\alpha + \beta > 1$, the firm has *increasing returns to scale*, and if $\alpha + \beta < 1$, it has *decreasing returns to scale*.

As an application, consider the carpet industry described in Example 6.4 (page 217). The production of both small and large firms can be described by Cobb-Douglas production functions. For small firms, $\alpha = .77$ and $\beta = .23$. Because $\alpha + \beta = 1$, there are constant returns to scale. For larger firms, however, $\alpha = .83$ and $\beta = .22$. Thus $\alpha + \beta = 1.05$, and there are increasing returns to scale. The Cobb-Douglas production function is frequently encountered in economics and can be used to model many kinds of production. We have already seen how it can accommodate differences in returns to scale. It can also account for changes in technology or productivity through changes in the value of A: The larger the value of A, more can be produced for a given level of K and L.

To find the amounts of capital and labor that the firm should utilize to minimize the cost of producing an output q_0, we first write the Lagrangian

$$\Phi = wL + rK - \lambda(AK^{\alpha}L^{\beta} - q_0) \qquad \textbf{(A7.15)}$$

Differentiating with respect to L, K, and λ, and setting those derivatives equal to 0, we obtain

$$\partial\Phi/\partial L = w - \lambda(\beta AK^{\alpha}L^{\beta-1}) = 0 \qquad \textbf{(A7.16)}$$

$$\partial\Phi/\partial K = r - \lambda(\alpha AK^{\alpha-1}L^{\beta}) = 0 \qquad \textbf{(A7.17)}$$

$$\partial\Phi/\partial\lambda = AK^{\alpha}L^{\beta} - q_0 = 0 \qquad \textbf{(A7.18)}$$

From equation (A7.16) we have

$$\lambda = w/A\beta K^{\alpha}L^{\beta-1} \qquad \textbf{(A7.19)}$$

Substituting this formula into equation (A7.17) gives us

$$r\beta AK^{\alpha}L^{\beta-1} = w\alpha AK^{\alpha-1}L^{\beta} \qquad \textbf{(A7.20)}$$

[2]For example, the marginal product of labor is given by $MP_L = \partial[F(K, L)]/\partial L = \beta AK^{\alpha}L^{\beta-1}$. Thus, MP_L falls as L increases.

or

$$L = \frac{\beta r}{\alpha w} K \tag{A7.21}$$

A7.21 is the expansion path. Now use Equation (A7.21) to substitute for L in equation (A7.18):

$$AK^{\alpha}\left(\frac{\beta r}{\alpha w} K\right)^{\beta} - q_0 = 0 \tag{A7.22}$$

We can rewrite the new equation as:

$$K^{\alpha+\beta} = \left(\frac{\alpha w}{\beta r}\right)^{\beta} \frac{q_0}{A} \tag{A7.23}$$

or

$$K = \left(\frac{\alpha w}{\beta r}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{q_0}{A}\right)^{\frac{1}{\alpha+\beta}} \tag{A7.24}$$

A7.24 is the factor demand for capital. We have now determined the cost-minimizing quantity of capital: Thus, if we wish to produce q_0 units of output at least cost, (A7.24) tells us how much capital we should employ as part of our production plan. To determine the cost-minimizing quantity of labor, we simply substitute equation (A7.24) into equation (A7.21):

$$L = \frac{\beta r}{\alpha w} K = \frac{\beta r}{\alpha w}\left[\left(\frac{\alpha w}{\beta r}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{q_0}{A}\right)^{\frac{1}{\alpha+\beta}}\right] \tag{A7.25}$$

$$L = \left(\frac{\beta r}{\alpha w}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{q_0}{A}\right)^{\frac{1}{\alpha+\beta}}$$

A7.25 is the constrained factor demand for labor. Note that if the wage rate w rises relative to the price of capital r, the firm will use more capital and less labor. Suppose that, because of technological change, A increases (so the firm can produce more output with the same inputs); in that case, both K and L will fall.

We have shown how cost-minimization subject to an output constraint can be used to determine the firm's optimal mix of capital and labor. Now we will determine the firm's cost function. The total cost of producing *any output* q can be obtained by substituting equations (A7.24) for K and (A7.25) for L into the equation $C = wL + rK$. After some algebraic manipulation we find that

$$C = w^{\beta/(\alpha+\beta)} r^{\alpha/(\alpha+\beta)}\left[\left(\frac{\alpha}{\beta}\right)^{\beta/(\alpha+\beta)} + \left(\frac{\alpha}{\beta}\right)^{-\alpha/(\alpha+\beta)}\right]\left(\frac{q}{A}\right)^{1/(\alpha+\beta)} \tag{A7.26}$$

This *cost function* tells us (1) how the total cost of production increases as the level of output q increases, and (2) how cost changes as input prices change. When $\alpha + \beta$ equals 1, equation (A7.26) simplifies to

$$C = w^{\beta} r^{\alpha}[(\alpha/\beta)^{\beta} + (\alpha/\beta)^{-\alpha}](1/A)q \qquad \textbf{(A7.27)}$$

In this case, therefore, cost will increase proportionately with output. As a result, the production process exhibits constant returns to scale. Likewise, if $\alpha + \beta$ is greater than 1, there are increasing returns to scale; if $\alpha + \beta$ is less than 1, there are decreasing returns to scale.

The firm's cost function contains many desirable features. To appreciate this fact, consider the special constant returns to scale cost function (A7.27). Suppose that we wish to produce q_0 in output but are faced with a doubling of the wage. How should we expect our costs to change? New costs are given by

$$C_1 = (2w)^{\beta} r^{\alpha}\left[\left(\frac{\alpha}{\beta}\right)^{\beta} + \left(\frac{\alpha}{\beta}\right)^{-\alpha}\right]\left(\frac{1}{A}\right)q_0 = 2^{\beta}\underbrace{w^{\beta} r^{\alpha}\left[\left(\frac{\alpha}{\beta}\right)^{\beta} + \left(\frac{\alpha}{\beta}\right)^{-\alpha}\right]\left(\frac{1}{A}\right)q_0}_{C_0} = 2^{\beta} C_0$$

Recall that at the beginning of this section, we assumed that $\alpha < 1$ and $\beta < 1$. Therefore, $C_1 < 2C_0$. Even though wages doubled, the cost of producing q_0 less than doubled. This is the expected result. If a firm suddenly had to pay more for labor, it would substitute away from labor and employ more of the relatively cheaper capital, thereby keeping the increase in total cost in check.

Now consider the dual problem of maximizing the output that can be produced with the expenditure of C_0 dollars. We leave it to you to work through this problem for the Cobb-Douglas production function. You should be able to show that equations (A7.24) and (A7.25) describe the cost-minimizing input choices. To get you started, note that the Lagrangian for this dual problem is $\Phi = AK^{\alpha}L^{\beta} - \mu(wL + rK - C_0)$.

EXERCISES

1. Of the following production functions, which exhibit increasing, constant, or decreasing returns to scale?
   a. $F(K, L) = K^2L$
   b. $F(K, L) = 10K + 5L$
   c. $F(K, L) = (KL)^{.5}$
2. The production function for a product is given by $q = 100KL$. If the price of capital is \$120 per day and the price of labor \$30 per day, what is the minimum cost of producing 1000 units of output?
3. Suppose a production function is given by $F(K, L) = KL^2$; the price of capital is \$10 and the price of labor \$15. What combination of labor and capital minimizes the cost of producing any given output?
4. Suppose the process of producing lightweight parkas by Polly's Parkas is described by the function

   $$q = 10K^{.8}(L - 40)^{.2}$$

   where q is the number of parkas produced, K the number of computerized stitching-machine hours, and L the number of person-hours of labor. In addition to capital and labor, \$10 worth of raw materials is used in the production of each parka.
   a. By minimizing cost subject to the production function, derive the cost-minimizing demands for K and L as a function of output (q), wage rates (w), and rental rates on machines (r). Use these results to derive the total cost function: that is, costs as a function of q, r, w, and the constant \$10 per unit materials cost.
   b. This process requires skilled workers, who earn \$32 per hour. The rental rate on the machines used in the process is \$64 per hour. At these factor prices, what are total costs as a function of q? Does this technology exhibit decreasing, constant, or increasing returns to scale?
   c. Polly's Parkas plans to produce 2000 parkas per week. At the factor prices given above, how many workers should the firm hire (at 40 hours per week) and how many machines should it rent (at 40 machine-hours per week)? What are the marginal and average costs at this level of production?