CHAPTER 8

Cost Curves

CHAPTER PREVIEW

The Chinese economy in the 1990s underwent an unprecedented boom. As part of that boom, enterprises such as HiSense Group grew rapidly.[1] HiSense, one of China's largest television producers, increased its rate of production by 50 percent per year during the mid-1990s. Its goal was to transform itself from a sleepy domestic producer of television sets into a consumer electronics giant whose brand name was recognized throughout Asia.

Of vital concern to HiSense and the thousands of other Chinese enterprises that were plotting similar growth strategies in the late 1990s was how production costs would change as its volume of output increased. There is little doubt that HiSense's production costs would go up as it produced more television sets. But *how fast* would they go up? HiSense's executives hoped that as it produced more television sets, the cost of *each television set* would go down, that is, its unit costs will fall as its annual rate of output goes up.

HiSense's executives also needed to know how input prices would affect its production costs. For example, HiSense competes with other large Chinese television manufacturers to buy up smaller factories. This competition bids up the price of capital. HiSense had to reckon with the impact of this price increase on its total production costs.

This chapter is about cost curves—relationships between costs and the volume of output. It picks up where Chapter 7 left off: with the comparative statics of the cost-minimization problem. The cost minimization-problem—both in the long run and the short run—gives rise to total, average, and marginal cost curves. This chapter studies these curves.

[1]This example is based on "Latest Merger Boom Is Happening in China and Bears Watching," *Wall Street Journal* (July 30, 1997), p. A1 and A9.

8.1 LONG-RUN COST CURVES

LONG-RUN TOTAL COST CURVES

In Chapter 7, we studied the firm's long-run cost minimization problem and saw how the cost-minimizing combination of labor and capital depended on the quantity of output Q and the prices of labor and capital, w and r. Figure 8.1(a) shows how the optimal input combination for a television firm, such as HiSense, changes as we vary output, holding input prices fixed. For example, when the firm produces 1 million televisions per year, the cost-minimizing input combination occurs at point A, with L_1 units of labor and K_1 units of capital. At this input combination, the firm is on an isocost line corresponding to TC_1 dollars of total cost, where $TC_1 = wL_1 + rK_1$. TC_1 is thus the minimized total cost when the firm pro-

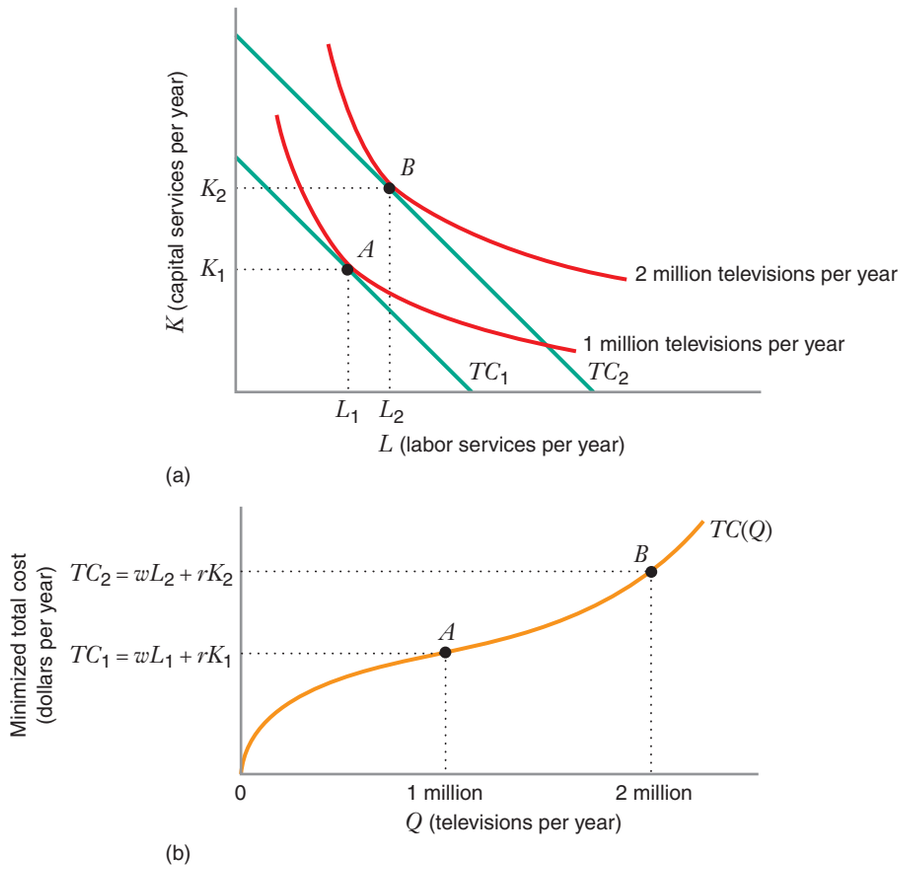

FIGURE 8.1 Cost Minimization and the Long-Run Total Cost Curve for a Producer of Television Sets
Panel (a) shows how the solution to the cost-minimization problem for a television producer changes as output changes from 1 million televisions per year to 2 million televisions per year. When output increases, the minimized total cost increases from TC_1 to TC_2. Panel (b) shows the long-run total cost curve. This curve shows the relationship between the volume of output and the minimum level of total cost the firm can attain when it produces that output.

duces 1 million units of output. As the firm increases output from 1 million to 2 million televisions per year, it ends up on an isocost line further out to point B, with L_2 units of labor and K_2 units of capital. Thus, its minimized total cost goes up (i.e., $TC_2 > TC_1$). It cannot be otherwise, because if the firm could decrease total cost by producing more output, it couldn't have been using a cost-minimizing combination of inputs in the first place.

Figure 8.1(b) shows the **long-run total cost curve,** denoted by $TC(Q)$. The long-run total cost curve shows how minimized total cost varies with output, holding input prices fixed. Because the cost-minimizing input combination moves us to higher isocost lines, the long-run total cost curve must be increasing in Q. We also know that when $Q = 0$, long-run total cost is 0. This is because, in the long run, the firm is free to vary all its inputs, and if it produces a zero quantity, the cost-minimizing input combination is zero labor and zero capital. Thus, comparative statics analysis of the cost-minimization problem implies that the *long-run total cost curve must be increasing and must equal* 0, *when* $Q = 0$.

LEARNING-BY-DOING EXERCISE 8.1

The Long-Run Total Cost Curve for a Cobb–Douglas Production Function

Let's return again to the production function $Q = 50L^{\frac{1}{2}}K^{\frac{1}{2}}$ that we analyzed in the Learning-By-Doing Exercises in Chapter 7.

Problem

(a) How does minimized total cost depend on the output Q and the input prices w and r for this production function?

Solution From Learning-By-Doing Exercise 7.4 in Chapter 7, we saw that the following equations described the cost-minimizing quantities of labor and capital:

$$L = \frac{Q}{50}\left(\frac{r}{w}\right)^{\frac{1}{2}} \tag{8.1}$$

$$K = \frac{Q}{50}\left(\frac{w}{r}\right)^{\frac{1}{2}} \tag{8.2}$$

To find the minimized total cost, we calculate the total cost the firm incurs when it uses this cost-minimizing input combination:

$$\begin{aligned} TC &= w\frac{Q}{50}\left(\frac{r}{w}\right)^{\frac{1}{2}} + r\frac{Q}{50}\left(\frac{w}{r}\right)^{\frac{1}{2}}, \\ &= \frac{Q}{50}w^{\frac{1}{2}}r^{\frac{1}{2}} + \frac{Q}{50}w^{\frac{1}{2}}r^{\frac{1}{2}} \\ &= \frac{w^{\frac{1}{2}}r^{\frac{1}{2}}}{25}Q. \end{aligned} \tag{8.3}$$

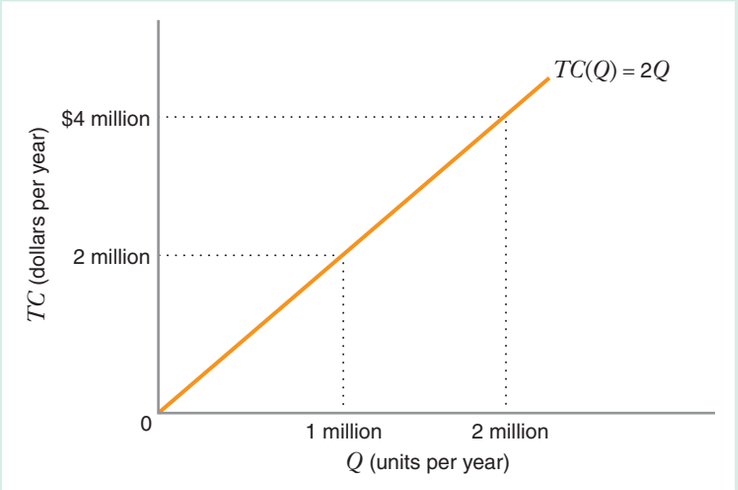

FIGURE 8.2 Long-Run Total Cost Curve for Learning-By-Doing Exercise 8.1
The long-run total cost curve for Learning-By-Doing Exercise 8.1 has the equation $TC(Q) = 2Q$.

Problem

(b) What is the graph of the long-run total cost curve when $w = 25$ and $r = 100$?

Solution Figure 8.2 shows that the graph of the long-run total cost curve is a straight line. We derive it by plugging $w = 25$ and $r = 100$ into expression (8.3) to get

$$TC(Q) = 2Q.$$

Similar Problem: 8.1, 8.3, 8.4

HOW DOES THE LONG-RUN TOTAL COST CURVE SHIFT WHEN INPUT PRICES CHANGE?

What Happens When Just One Input Price Changes?

In the introduction, we discussed how HiSense faced the prospect of higher prices for certain inputs, such as capital. To illustrate how an increase in an input price affects a firm's total cost curve, let's return to the cost-minimization problem for our hypothetical television producer. Figure 8.3 shows what happens when the price of capital increases, holding output and the price of labor constant. Suppose that at the initial situation, the optimal input combination for an annual output of 1 million television sets occurs at point A, and the minimized total cost is $50 million per year. The figure shows that after the increase in the price of capital, the optimal input combination, point B, must lie along an isocost line corresponding to a total cost that is *greater* than $50 million. To see why, note that

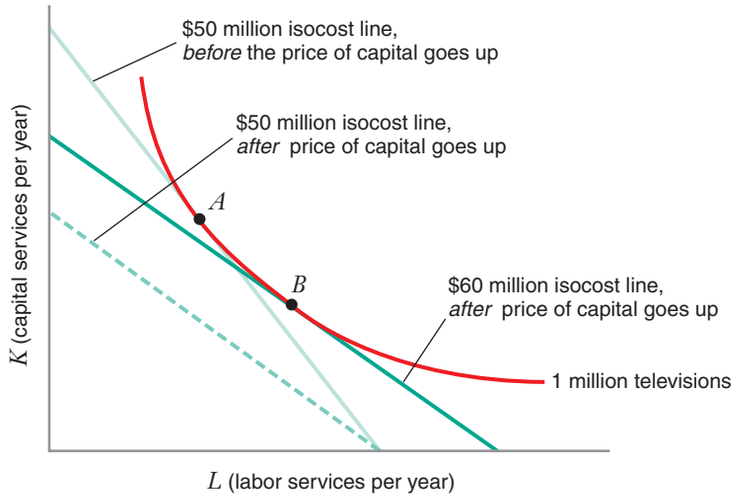

FIGURE 8.3 How a Change in the Price of Capital Affects the Optimal Input Combination and Long-Run Total Cost for a Producer of Television Sets
Initially, the optimal input combination is point A, and the minimized total cost is $50 million. After the price of capital goes up, the optimal input combination B lies on an isocost corresponding to a higher level of cost, $60 million. The increase in the price of labor thus increases the firm's long-run total cost.

the $50 million isocost line *at the new input prices* intersects the horizontal axis in the same place as the $50 million isocost line *at the old input prices*. However, the new $50 million isocost line is flatter because the price of capital has gone up. You can see from Figure 8.3 that the firm could not operate on the $50 million isocost line because it would be unable to produce the desired quantity of 1 million television sets. To produce 1 million television sets, the firm must operate on an isocost line that is further to the northeast and thus corresponds to a higher level of cost ($60 million perhaps). Thus, holding output fixed, the minimized total cost goes up when the price of capital goes up.[2]

This analysis then implies that an increase in the price of capital results in a new total cost curve that lies above the original total cost curve at every $Q > 0$. At $Q = 0$, long-run total cost is still zero. Thus, as Figure 8.4 shows, an increase in an input price rotates the long-run total cost curve upward.[3]

[2]An analogous argument would show that minimized total cost would go down when the price of capital goes down.

[3]There is one case in which an increase in an input price would not affect the long-run total cost curve. If the firm is initially at a corner point solution using a zero quantity of the input, an increase in the price of the input will leave the firm's cost-minimizing input combination—and thus its minimized total cost—unchanged. In this case, the increase in the input price would not shift the long-run total cost curve.

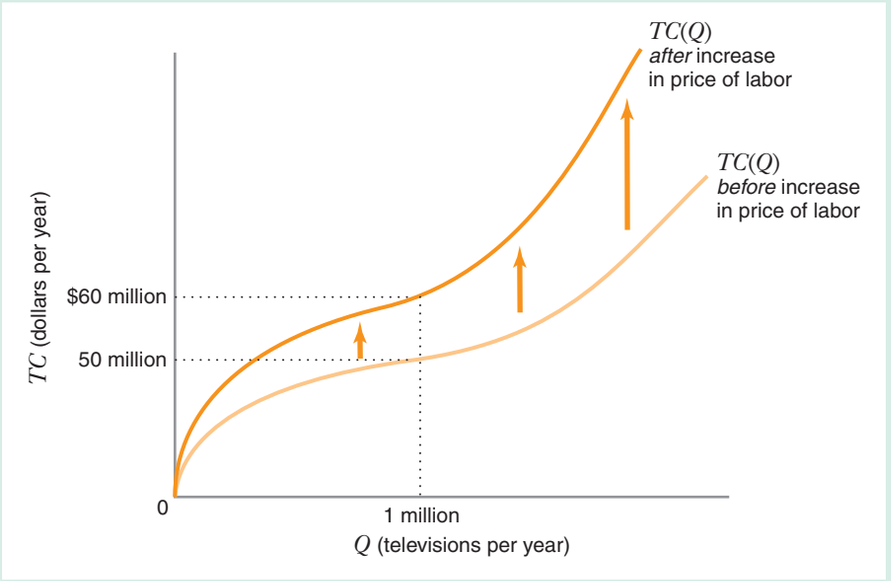

FIGURE 8.4 How a Change in the Price of Capital Affects the Long-Run Total Cost Curve for a Producer of Television Sets
An increase in the price of capital results in a new long-run total cost curve that lies above the initial long-run total cost curve at every quantity except $Q = 0$. For example, at the quantity of 1 million units per year, long-run total cost increases from $50 million to $60 million per year. Thus, the increase in the price of capital rotates the long-run total cost curve upward.

What Happens to Long-Run Total Cost When All Input Prices Change Proportionately?

What if the price of capital and the price of labor both go up by the same percentage amount, say 10 percent? Returning once again to the cost-minimization problem, we see from Figure 8.5 that *a proportionate increase in both input prices leaves the optimal input combination unchanged*. The slope of the isocost line stays the same because it equals the ratio of the price of labor to the price of capital. Because both input prices increased by the same percentage amount, this ratio remains unchanged.

However, the total cost curve must shift in a special way. Since the optimal input combination remains the same, a 10 percent increase in the prices of all inputs must increase the minimized total cost by exactly 10 percent! More generally, any given percentage increase in *all* input prices will do the following:

- Leave the optimal input combination unchanged, *and*
- Shift up the total cost curve by exactly the same percentage as the common increase in input prices.

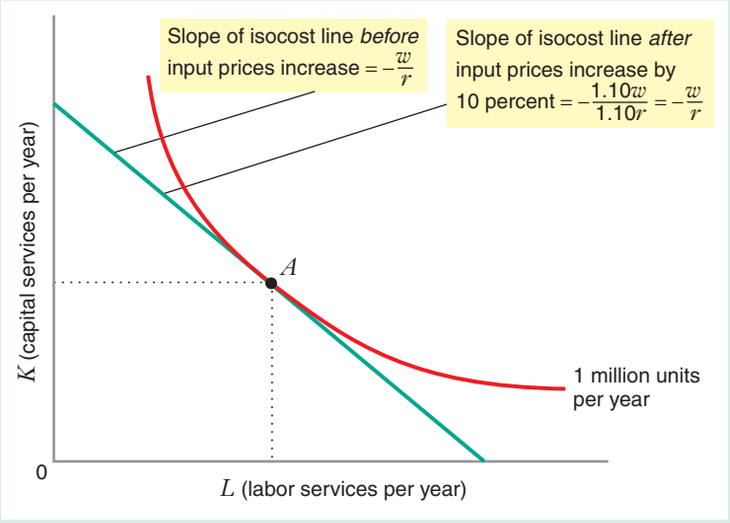

FIGURE 8.5 How a Proportionate Change in the Prices of All Inputs Affects the Cost-Minimizing Input Combination
A 10 percent increase in the prices of all inputs leaves the slopes of the isocost lines unchanged. Thus, the cost-minimizing input combination for a particular output level, such as 1 million units, remains the same.

EXAMPLE 8.1

How Would Input Prices Affect the Long-Run Total Costs for a Trucking Firm?[4]

The intercity trucking business is a good setting in which to study the behavior of long-run total costs because when input prices or output changes, trucking firms can adjust their input mixes without too much difficulty. Drivers can be hired or laid off easily, and trucks can be bought or sold as circumstances dictate. There is also considerable data on output, expenditures on inputs, and input quantities, so we can use statistical techniques to estimate how total cost varies with input prices and output. Utilizing such data, Richard Spady and Ann Friedlaender estimated long-run total cost curves for trucking firms that carry general merchandise. Many semis fall into this category.

Trucking firms use three major inputs: labor, capital (e.g., trucks), and diesel fuel. Their output is transportation services, usually measured as ton-miles per year. One ton-mile is one ton of freight carried one mile. A trucking company that hauls 50,000 tons of freight 100,000 miles during a given year would thus have a total output of 50,000 $\times$ 100,000, or 5,000,000,000 ton-miles per year.

Figure 8.6 illustrates an example of the cost curve estimated by Spady and Friedlaender. Note that total cost increases with the quantity of output, as the theory we just discussed implies. Total cost also increases in the prices of inputs. Figure 8.6 shows how doubling the price of labor (holding all other input prices fixed) affects the total cost curve. The increase in the input price shifts the total cost curve upward at every point except $Q = 0$. Figure 8.6 also shows the effect of doubling the price of capital and doubling the price of fuel. These increases also shift the total cost curve upward, though this shift is not as much as when the price of labor goes up. This analysis shows that the total cost of a trucking firm is most sensitive to changes in the price of labor and least sensitive to changes in the price of diesel fuel. ■

[4]This example draws from A. F. Friedlaender, and R. H. Spady, *Freight Transport Regulation: Equity, Efficiency, and Competition in the Rail and Trucking Industries* (Cambridge, MA: MIT Press, 1981).

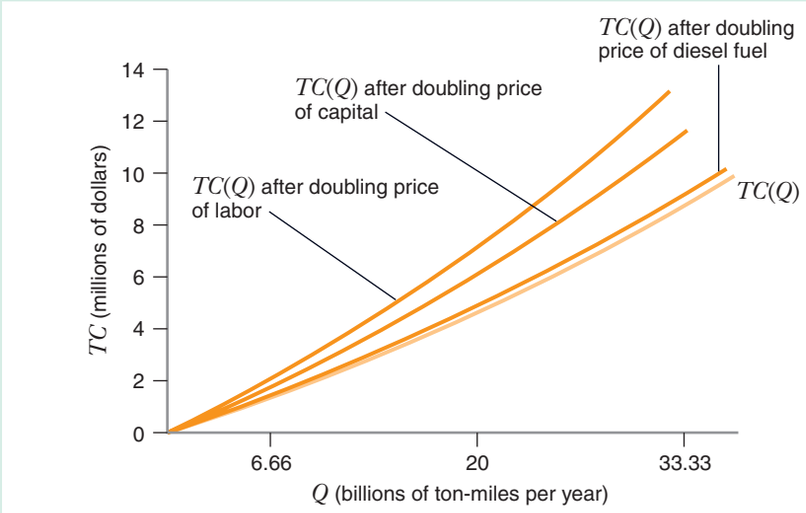

FIGURE 8.6 Long-Run Total Cost Curve for a Trucking Firm
The curve $TC(Q)$ is a graph of the long-run total cost function for a typical trucking firm. Doubling the price of labor shifts the long-run total cost function upward, as does doubling the prices of capital and diesel fuel. However, an increase in the price of labor has a bigger impact on total cost than either an increase in the price of capital or diesel fuel.

8.2 LONG-RUN AVERAGE AND MARGINAL COST

WHAT ARE LONG-RUN AVERAGE AND MARGINAL COSTS?

Two other types of cost play an important role in microeconomics: long-run average cost and long-run marginal cost. **Long-run average cost** is the firm's cost per unit of output. It equals long-run total cost divided by Q:

$$AC(Q) = \frac{TC(Q)}{Q}.$$

Long-run marginal cost is the rate of change at which long-run total cost changes with respect to output:

$$MC(Q) = \frac{TC(Q + \Delta Q) - TC(Q)}{\Delta Q}$$

$$= \frac{\Delta TC}{\Delta Q}.$$

Although long-run average and marginal cost are both derived from the firm's long-run total cost curve, the two costs are generally different. Average cost is the cost per unit that the firm incurs in producing all of its output. Marginal cost, by contrast, is the increase in cost from producing an additional unit of output.

Figure 8.7 illustrates the difference between marginal and average cost. At a particular output level, such as 50 units per year, average cost is equal to the slope of ray 0*A*. This slope is equal to \$1,500/50 units, so the firm's average cost when it produces 50 units per year is \$30 per unit. By contrast, the marginal cost when the firm produces 50 units per year is the slope of the total cost curve at a quantity of 50. In Figure 8.7 this is represented by the slope of the line *BAC* that is tangent to the total cost curve at a quantity of 50 units. The slope of this tangent line is 10, so the firm's marginal cost at a quantity of 50 units is \$10 per unit. As we vary total output, we can trace out the long-run average cost curve by imagining how the slope of rays such as 0*A* change as we move along the long-run total cost curve. Similarly, we can trace out the long-run marginal cost curve by imagining how the slope of tangent lines such as *BAC* change as we move along the total cost curve. As Figure 8.7 shows, these two "thought processes" will generate two different curves.

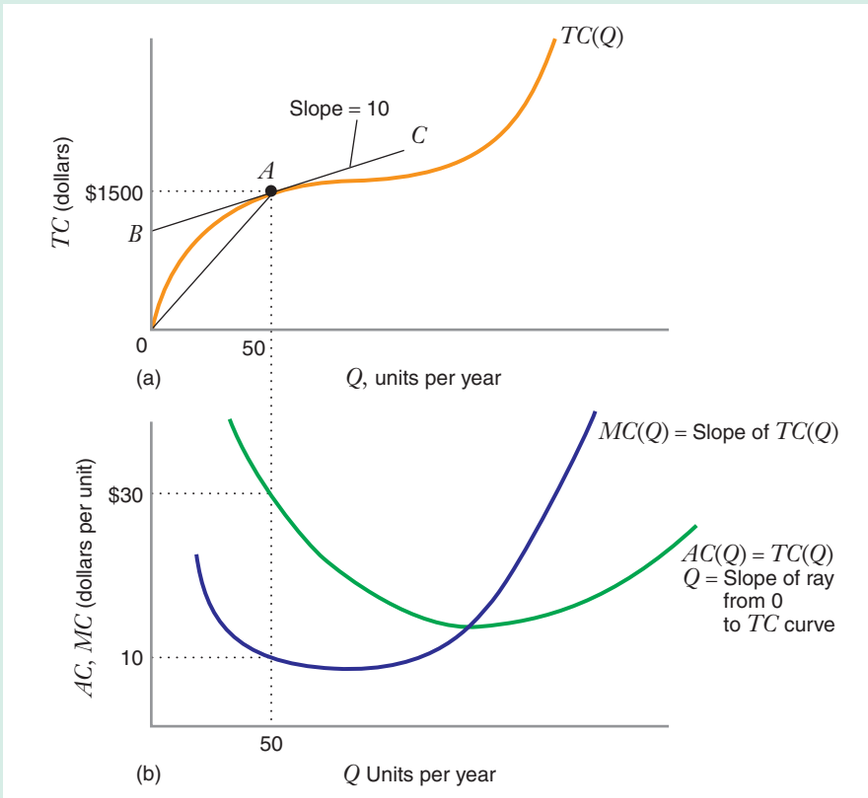

FIGURE 8.7 Deriving Average and Marginal Cost from the Total Cost Curve
The top panel shows the firm's total cost curve. The average cost at a quantity of 50 units, is the slope of the ray from 0*A*, or \$30 per unit. The marginal cost at a quantity of 50 units is the slope of the total cost curve at this quantity, which equals the slope of tangent line *BAC*. This tangent line's slope is 10, so marginal cost at 50 units is \$10 per unit. More generally, we can trace out the average cost curve by imagining how the slope of rays from 0 to the total cost curve (such as 0*A*) change as we move along the the total cost curve. We can trace out the marginal cost curve by imagining how the slope of tangent line (such as *BAC*) change as we move along the total cost curve.

LEARNING-BY-DOING EXERCISE 8.2

Deriving Long-Run Average and Marginal Costs from a Long-Run Total Cost Curve

Average cost and marginal cost are often different. However, there is one special case in which they are the same.

Problem In Learning-By-Doing Exercise 8.1 we derived the long-run total cost curve for a Cobb–Douglas production function. For particular input prices ($w = 25$ and $r = 100$), the long-run total cost curve was described by the equation $TC(Q) = 2Q$. What are the long-run average and marginal cost curves associated with this long-run total cost curves?

Solution Long-run average cost is

$$AC(Q) = \frac{2Q}{Q} = 2.$$

Note that average cost does not depend on Q. Its graph would be a horizontal line, as Figure 8.8 shows.

Long-run marginal cost is

$$MC(Q) = \frac{\Delta(2Q)}{\Delta Q} = 2.$$

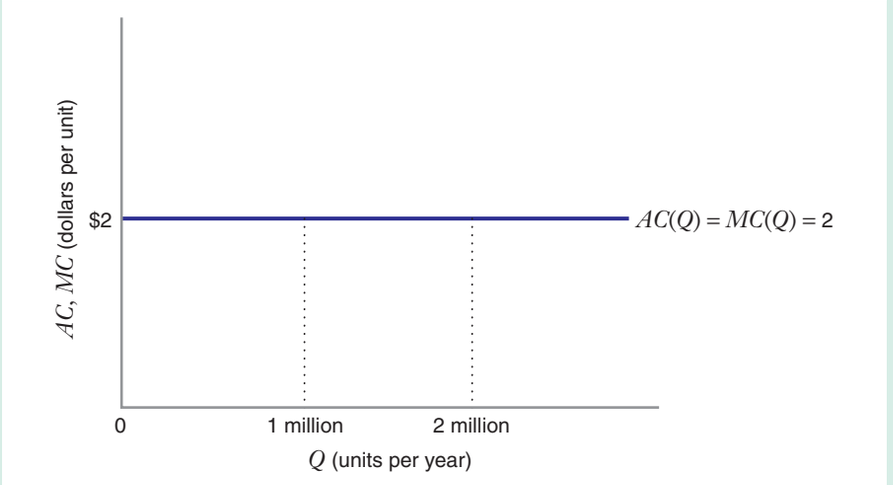

FIGURE 8.8 Long-Run Average and Marginal Cost Curves for Learning-By-Doing Exercise 8.2
The long-run average and marginal cost curves in Learning-By-Doing Exercise 8.2 are identical horizontal lines.

Long-run marginal cost also does not depend on Q. In fact, it is identical to the long-run average cost curve, so its graph is also a horizontal line.

This exercise illustrates a general point. Whenever the long-run total cost is a straight line (as in Figure 8.2), long-run average and long-run marginal cost will be the same, and their common graph will be a horizontal line.

Similar Problem: 8.2

RELATIONSHIP BETWEEN LONG-RUN MARGINAL AND AVERAGE COST CURVES

As with other average and marginal concepts you will study in this book (e.g., average product versus marginal product), there is a systematic relationship between the long-run average and long-run marginal cost curves. Figure 8.9 illustrates this relationship:

- When marginal cost is *less than* average cost, average cost is *decreasing in quantity*. That is, if $MC(Q) < AC(Q)$, $AC(Q)$ decreases in Q.
- When marginal cost is *greater than* average cost, average cost is *increasing in quantity*. Thus is, if $MC(Q) > AC(Q)$, $AC(Q)$ increases in Q.
- When marginal cost is *equal to* average cost, average cost *neither increases nor decreases in quantity*. Either its graph is flat, or we are at a point at which $AC(Q)$ is minimized in Q.

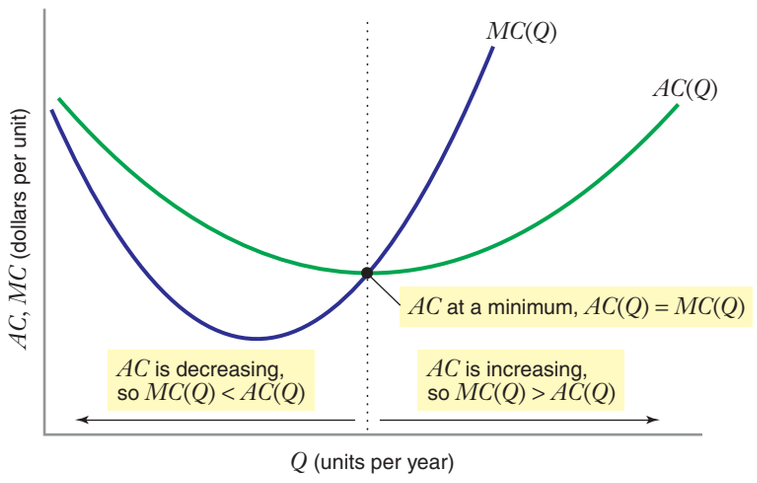

FIGURE 8.9 Relationship Between the Average and Marginal Cost Curves
When average cost is decreasing, marginal cost is less than average cost. When average cost is increasing, marginal cost is greater than average cost. When average cost attains its minimum, marginal cost equals average cost.

The relationship between marginal cost and average cost is the same as the relationship between the marginal of anything and the average of anything. To illustrate this point, suppose that the average height of students in your class is 160 cm. Now, a new student, Mike Margin, joins the class, and the average height rises to 161 cm. What do we know about his height? Since the average height is increasing, the "marginal height" (Mike Margin's height) must be above the average. If the average height had fallen to 159 cm, it would have been because his height was below the average. Finally, if the average height had remained the same when Mr. Margin joined the class, his height had to exactly equal the average height in the class.

The relationship between average and marginal height in your class is the same as the relationship between average and marginal product that we observed in Chapter 6. It is also the relationship between average and marginal cost that we just described. And it is the relationship between average and marginal revenue that we will study in Chapter 11.

EXAMPLE 8.2 *The Relationship Between Average and Marginal Cost in Higher Education*

How big is your college or university? Is it a large school, such as Ohio State, or a smaller university, such as Northwestern? At which school is the cost per student likely to be lower? Does university size affect the average and marginal cost of "producing" education?

Rajindar and Manjulika Koshal recently studied how size affects the average and marginal cost of education.[5] They collected data on the average cost per student from 195 U.S. universities from 1990 to 1991 and estimated an average cost curve for these universities.[6] To control for differences in cost that stem from differences among universities in terms of their commitment to graduate programs, the Koshals estimated average cost curves for four groups of universities, primarily distinguished by the number of Ph.Ds awarded per year and the amount of government funding for Ph.D. students these universities received. For simplicity, we discuss the cost curves for the category that includes the 66 universities nationwide with the largest graduate programs (e.g., schools like Harvard, Ohio State, Northwestern, and the University of California at Berkeley).

Figure 8.10 shows the estimated average and marginal cost curves for this category of schools. It shows that the average cost per student declines until about

[5]R. Koshal and M. Koshal, "Quality and Economies of Scale in Higher Education," *Applied Economics 27* (1995): 773–778.

[6]To control for variations in cost that might be due to differences in academic quality, their analysis also allowed average cost to depend on the student-faculty ratio and the academic reputation of the school, as measured by factors, such as average SAT scores of entering freshmen. In the graph in Figure 8.10, these variables are assumed to be equal to their national averages.

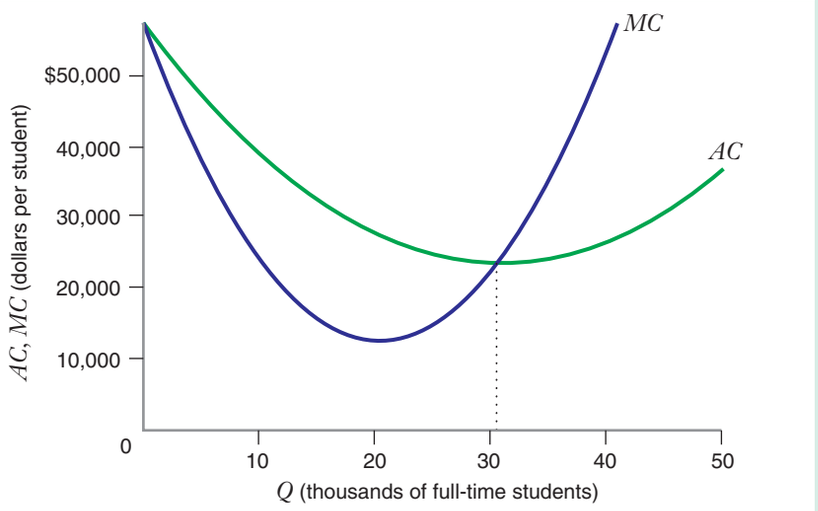

FIGURE 8.10 The Average and Marginal Cost Curves for University Education at U.S. Universities
The marginal cost of an additional student is less than the average cost per student until enrollment reaches about 30,000 students. Until that point, average cost per student falls with the number of students. Beyond that point, the marginal cost of an additional student exceeds the average cost per student, and average cost increases with the number of students.

30,000 full-time undergraduate students (about the size of Indiana University, for example). Because few universities are this large, the Koshals' research suggests that for most universities in the United States with large graduate programs, the marginal cost of an additional undergraduate student is less than the average cost per student, and thus an increase in the size of the undergraduate student body would reduce the cost per student.

This finding seems to make sense. Think about your university. It already has a library and buildings for classrooms. It already has a president and a staff to run the school. These costs will probably not go up much if more students are added. Adding additional students is, of course, not costless. For example, more classes might have to be added. But it is not *that* difficult to find people who are able and willing to teach university classes (e.g., graduate students). Until the point is reached at which more dormitories or additional classrooms are needed, the extra costs of more students are not likely to be that large. Thus, for the typical university, while the *average* cost per student might be fairly high, the *marginal* cost of matriculating an additional student is often fairly low. If so, average cost will decrease with the number of students. ■

ECONOMIES AND DISECONOMIES OF SCALE

The term **economies of scale** describes a situation in which average cost decreases as output goes up, and **diseconomies of scale** describes the opposite: average cost increases as output goes up. Economies and diseconomies of scale are important concepts. The extent of economies of scale can affect the structure of an industry. Economies of scale can also explain why some firms are more profitable than others in the same industry. Claims of economies of scale are often used to justify mergers between two firms producing the same product.[7]

Figure 8.11 illustrates economies and diseconomies of scale by showing an average cost curve that many economists believe typifies real-world production processes. For this average cost curve, there is an initial range of economies of scale (0 to Q'), followed by a range over which average cost is flat (Q' to Q''), and eventually a range of diseconomies of scale ($Q > Q''$).

Economies of scale have various causes. They may result from the physical properties of processing units that give rise to increasing returns to scale in inputs (e.g., the cube-square rule discussed in Chapter 6). Economies of scale can also arise due to specialization of labor. As the number of workers increases with the output of the firm, workers can specialize on tasks, which often increases their productivity. Specialization can also eliminate time-consuming changeovers of workers and equipment. This too would increase worker productivity and lower unit costs.

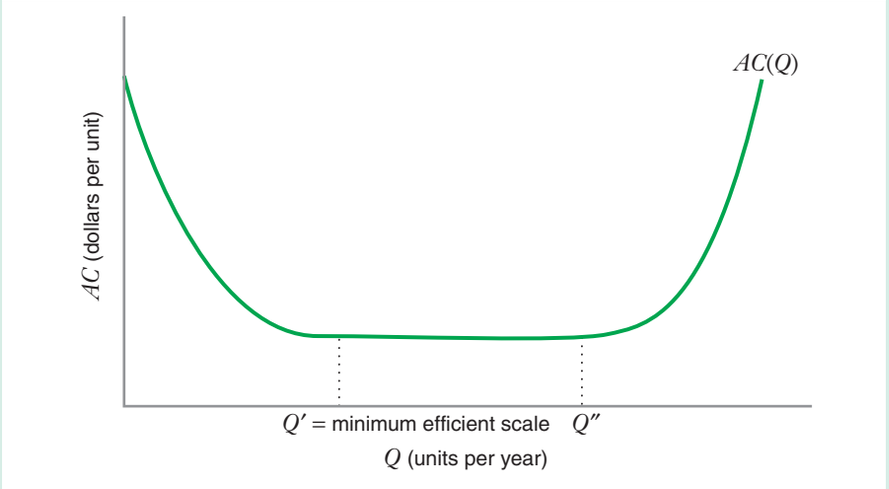

FIGURE 8.11 Real-World Average Cost Curve
This average cost curve typifies many real-world production processes. There are economies of scale for outputs less than Q'. Average costs are flat between Q' and Q'', and there are diseconomies of scale thereafter. The output level Q' at which the economies of scale are exhausted is called the minimum efficient scale.

[7]See Chapter 4 of F. M. Scherer and D. Ross, *Industrial Market Structure and Economic Performance* (Boston: Houghton Mifflin) 1990, for a detailed discussion of the implications of economies of scale for market structure and firm performance.

Economies of scale may also result from the need to employ **indivisible inputs.** An indivisible input is an input that is available only in a certain minimum size; its quantity cannot be scaled down as the firm's output goes to zero. An example of an indivisible input is a high-speed packaging line for breakfast cereal. Even the smallest such lines have huge capacity, 14 million pounds of cereal per year. A firm that might only want to produce 5 million pounds of cereal a year would still have to purchase the services of this indivisible piece of equipment.

Indivisible inputs lead to decreasing average costs (at least over a certain range of output) because when a firm purchases the services of an indivisible input, it can "spread" the cost of the indivisible input over more units of output as output goes up. For example, a firm that purchases the services of a minimum-scale packaging line to produce 5 million pounds of cereal per year will incur the same total cost on this input when it increases production to 10 million pounds of cereal per year.[8] This will drive the firm's average costs down.

The region of diseconomies of scale in Figure 8.11 is usually thought to occur because of **managerial diseconomies.** Managerial diseconomies arise when a given percentage increase in output forces the firm to increase its spending on the services of managers by more than this percentage. To see why managerial diseconomies of scale can arise, imagine an enterprise whose success depends on the talents or insight of one key individual (e.g., the entrepreneur who started the business). As the enterprise grows, that key individual cannot be replicated. To compensate, the firm may have to employ enough additional managers that total costs increase at a faster rate than output, which then pushes average costs up. Viewed this way, managerial diseconomies are another example of diminishing marginal returns to variable inputs that arise when certain other inputs (specialized managerial talent) are in fixed supply.

The smallest quantity at which the long-run average cost curve attains its minimum point is called the **minimum efficient scale,** or **MES.** The MES occurs at output Q' in Figure 8.11. The magnitude of MES relative to the size of the market often indicates the magnitude of economies of scale in particular industries. The larger is MES in comparison to overall market sales, the greater the magnitude of economies of scale. Table 8.1 shows MES as a percentage of total industry output, for a selected group of U.S. food and beverage industries.[9] The industries with the largest MES-market size ratios are breakfast cereal and cane sugar refining. These industries have significant economies of scale. The industries with the lowest MES-market size ratios are mineral water and bread. Economies of scale in manufacturing in these industries appear to be weak.

[8]Of course, it may spend more on other inputs, such as raw materials, that are not indivisible.

[9]In this table, MES is measured as the capacity of the median plant in an industry. The median plant is the plant whose capacity lies exactly in the middle of the range of capacities of plants in an industry. That is, 50 percent of all plants in particular industry have capacities that are smaller than the median plant in that industry, and 50 percent have capacities that are larger. Estimates of MES based on the capacity of the median plant correlate highly with "engineering estimates" of MES that are obtained by asking well-informed manufacturing and engineering personnel to provide educated estimates of minimum efficient scale plant sizes. Data on median plant size in U.S. industries are available from the U.S. Census of Manufacturing.

TABLE 8.1
MES as a Percentage of Industry Output for Selected U.S. Food and Beverage Industries

Industry	MES as % of Output	Industry	MES as % of Output
Beet sugar	1.87	Breakfast cereal	9.47
Cane sugar	12.01	Mineral water	0.08
Flour	0.68	Roasted coffee	5.82
Bread	0.12	Pet food	3.02
Canned vegetables	0.17	Baby food	2.59
Frozen food	0.92	Beer	1.37
Margarine	1.75		

Source: Table 4.2 in J. Sutton, *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration* (Cambridge, MA: MIT Press, 1991).

EXAMPLE 8.3

Economies of Scale in Alumina Refining[10]

Manufacturing aluminum involves several steps, one of which is alumina refining. Alumina is a chemical compound consisting of aluminum and oxygen atoms (Al_2O_3). Alumina is created when bauxite ore—the basic raw material used to produce aluminum—is transformed using a technology known as the Bayer process.

There are substantial economies of scale in the refining of alumina. Table 8.2—drawn from John Stuckey's study of the aluminum industry—shows estimated long-run average costs as a function of the capacity of an alumina refinery. As plant capacity doubles from 150,000 tons per year to 300,000 tons per year, long-run average cost declines by about 12 percent. Stuckey reports that average costs in alumina refining may continue to fall up to capacities of 500,000. If so, then the minimum efficient scale of an alumina refinery would occur at an output of 500,000 tons per year.

If firms understand this, we would expect most alumina plants to have capacities of at least 500,000 tons per year. In fact, this is true. In 1979, the average capacity of the 10 alumina refineries in North America was 800,000 tons per year, and only two were under 500,000 tons per year. No alumina refinery's capacity exceeded 1.3 million tons per year. This suggests that diseconomies of scale set in at about this level of output. ■

[10]The information in this example draws from J. Stuckey, *Vertical Integration and Joint Ventures in the Aluminum Industry* (Cambridge, MA: Harvard University Press, 1983), especially pp. 12–14.

TABLE 8.2
Plant Capacity and Average Cost in Alumina Refining

Plant Capacity (tons)	Index of Average Cost (equals 100 at 300,000 tons)
55,000	139
90,000	124
150,000	114
300,000	100

Source: Table 1-1 in Stuckey, *Vertical Integration and Joint Ventures in the Aluminum Industry.* (Cambridge, MA: Harvard University Press, 1983.)

Economies of Scale for "Backoffice" Activities in a Hospital

EXAMPLE 8.4

The business of health care was in the news a lot during the 1990s. One of the most interesting trends was the consolidation of hospitals through mergers. In the Chicago area, for example, Northwestern Memorial Hospital merged with several suburban hospitals, such as Evanston Hospital, to form a large multi-hospital system covering the North Side of Chicago and the North Shore.

Proponents of hospital mergers argue that mergers enable hospitals to achieve cost savings through economies of scale in "backoffice" operations—activities, such as laundry, housekeeping, cafeterias, printing and duplicating services, and data processing, that do not generate revenue for a hospital directly, but that the hospital cannot function without. Opponents argue that such cost savings are illusory and that hospital mergers mainly reduce competition in local hospital markets. The U.S. antitrust authorities have blocked several hospital mergers on this basis.

David Dranove recently studied the extent to which backoffice activities within a hospital are subject to economies of scale.[11] Figure 8.12 summarizes some of his findings. The figure shows the average cost curves for three different activities: cafeterias, printing and duplicating, and data processing. Output is measured as the annual number of patients who are discharged by the hospital. (For each activity, average cost is normalized to equal $1, at an output of 10,000 patients per year.) These figures show that economies of scale vary from activity to activity. Cafeterias are characterized by significant economies of scale. For printing and duplicating, the average cost curve is essentially flat. And for data processing, diseconomies of scale arise at a fairly low level of output. Overall, averaging the 14 backoffice activities that he studied, Dranove found that there are economies of scale in these activities, but they are largely exhausted at an output of about 7,500 patient discharges per year. This would correspond to a hospital with 200 beds, which is medium-sized by today's standards.

Dranove's analysis shows that a merger of two large hospitals would be unlikely to achieve additional economies of scale in backoffice operations. This suggests that claims that hospital mergers generally reduce costs per patient should be viewed with skepticism, unless both merging hospitals are small. ■

[11]"Economies of Scale in Nonrevenue Producing Cost Centers: Implications for Hospital Mergers," working paper, Northwestern University, February 1997.

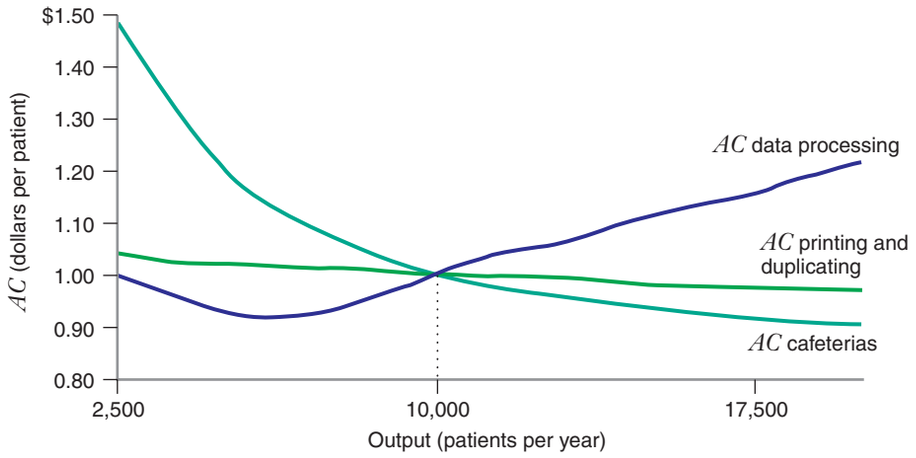

FIGURE 8.12 Average Cost Curves for "Back-office" Activities in a Hospital
The figure shows average cost curves for three "back-office" activities in a hospital: cafeterias, printing and duplicating, and data processing. Cafeterias exhibit significant economies of scale. Data processing exhibits diseconomies of scale beyond an output of about 5,000 patients per year. And the average cost curve for printing and duplicating is essentially flat, so that there are no significant economies or diseconomies of scale in this activity.

RETURNS TO SCALE VERSUS ECONOMIES OF SCALE

The concept of economies of scale is closely related to the concept of returns to scale introduced in Chapter 6. The returns to scale of the production function will determine how average cost varies with output and thus the existence of economies or diseconomies of scale.

We can illustrate this point most clearly with a single-input production function. Table 8.3 shows three different production functions in which output Q is a function of the quantity of labor L. The first exhibits constant returns to scale (CRTS); the second exhibits increasing returns to scale (IRTS), and the third

TABLE 8.3
Relationship Between Returns to Scale and the Long-Run Average Cost Curve

	CRTS	IRTS	DRTS
Production Function	$Q = L$	$Q = L^2$	$Q = \sqrt{L}$
Labor Requirements Function	$L = Q$	$L = \sqrt{Q}$	$L = Q^2$
Total Cost	$TC = wQ$	$TC = w\sqrt{Q}$	$TC = wQ^2$
Average Cost	$AC = w$	$AC = \frac{w}{\sqrt{Q}}$	$AC = wQ$
How Does AC Vary with Q?	Constant	Decreasing	Increasing

exhibits decreasing returns to scale (DRTS). Table 8.3 also shows the labor-requirements functions for these three production functions.[12] It also shows expressions for total cost and average cost, given a price of labor w. For the production function exhibiting constant returns to scale, the average cost function is independent of the quantity of output (i.e., it equals w no matter what Q is). For the production function exhibiting increasing returns to scale, average cost is a decreasing function of the quantity of output Q (i.e., as Q goes up, AC goes down). And for the production function exhibiting decreasing returns to scale, the average cost is an increasing function of output (i.e., as Q goes up, AC also goes up).

This can be summarized in three general relationships:

- When the production function exhibits *increasing returns to scale*, the long-run average cost curve exhibits *economies of scale* (i.e., $AC(Q)$ must decrease in Q).
- When the production function exhibits *decreasing returns to scale*, the long-run average cost curve exhibits *diseconomies of scale* (i.e., $AC(Q)$ must increase in Q).
- When the production function exhibits *constant returns to scale*, the long-run average cost curve is flat: It neither increases nor decreases in output.

MEASURING THE EXTENT OF ECONOMIES OF SCALE: THE OUTPUT ELASTICITY OF TOTAL COST

In Chapter 2, you learned that elasticities of demand, such as the price elasticity of demand or income elasticity of demand, tell us how sensitive demand is to the various factors that drive demand, such as price or income. We can also use elasticities to tell us how sensitive total cost is to the factors that influence it. An important cost elasticity is the **output elasticity of total cost,** denoted by $\epsilon_{TC,Q}$. It is defined as the percentage change in total cost per 1 percent change in output:

$$\epsilon_{TC,Q} = \frac{\frac{\Delta TC}{TC}}{\frac{\Delta Q}{Q}}.$$

We can rewrite this as follows:

$$\epsilon_{TC,Q} = \frac{\Delta TC}{\Delta Q} \div \frac{TC}{Q} = \frac{MC}{AC}.$$

Because the output elasticity of total cost is equal to the ratio of marginal to average cost, it tells us whether there are economies of scale or diseconomies of scale. This is because the following conditions hold:

- If $\epsilon_{TC,Q} < 1$, $MC < AC$, so AC decreases in Q, and we have *economies of scale*.
- If $\epsilon_{TC,Q} > 1$, $MC < AC$, so AC increases in Q, and we have *diseconomies of scale.*
- If $\epsilon_{TC,Q} = 1$, $MC = AC$, so AC neither increases nor decreases in Q.

The output elasticity is often used to characterize the nature of economies of scale in different industries. Table 8.4, for example, shows results of a recent study that estimated the output elasticity of total cost for several manufacturing industries

[12]Recall from Chapter 6 that the labor requirements function tells us the quantity of labor needed to produce a given amount of output.

TABLE 8.4
Estimates of the Output Elasticities for Selected Manufacturing Industries in India

Industry	Output Elasticity of Total Cost
Iron and Steel	0.553
Cotton Textiles	1.211
Cement	1.162
Electricity and Gas	0.3823

in India.[13] Iron and steel industries and electricity and gas industries have output elasticities significantly less than 1, indicating the presence of economies of scale. By contrast, textile and cement firms' output elasticities are a little higher than 1, indicating slight diseconomies of scale.[14]

8.3 SHORT-RUN COST CURVES

The long-run total cost curve shows how the firm's minimized total cost varies with output when the firm is free to adjust all its inputs. The **short-run total cost curve**, $STC(Q)$, tells us the minimized total cost of producing Q units of output when capital is fixed at a particular level, $\overline{K}$.. The short-run total cost curve is the sum of two components: the **total variable cost curve,** $TVC(Q)$, and the **total fixed cost curve** TFC (i.e., $STC(Q) = TVC(Q) + TFC$). The total variable cost curve $TVC(Q)$ is the sum of expenditures on variable inputs, such as labor and materials, at the short-run cost-minimizing input combination. Total fixed cost is equal to the cost of the fixed capital services (i.e., $TFC = r\overline{K}$) and thus does not vary with output. Figure 8.13 shows a graph of the short-run total cost curve, the total variable cost curve, and the total fixed cost curve.

LEARNING-BY-DOING EXERCISE 8.3

Deriving the Short-Run Total Cost Curve

Let us return to the production function in Learning-By-Doing Exercise 7.6 in Chapter 7. For that production function, the firm uses three inputs: capital, labor, and materials:

$$Q = K^{\frac{1}{2}}L^{\frac{1}{4}}M^{\frac{1}{4}}$$

[13]R. Jha, M.N. Murty, S. Paul, and B. Bhaskara Rao, "An Analysis of Technological Change, Factor Substitution, and Economies of Scale in Manufacturing Industries in India, *Applied Economics 25* (October 1993): 1337–1343. The estimated output elasticities are reported in Table 5.

[14]The estimated output elasticities for textiles and cement are not *statistically* different from 1. Thus, these industries might be characterized by constant returns to scale.

Problem What is the short-run total cost curve for this production function when capital is fixed at a level $\overline{K}$ and input prices are $w = 16$, $m = 1$, and $r = 2$? What are total variable cost and total fixed cost?

Solution In Learning-By-Doing Exercise 7.6, we derived the short-run cost-minimizing quantities are labor and materials for this production function:

$$L = \frac{Q^2}{4\overline{K}}$$

$$M = \frac{4Q^2}{\overline{K}}$$

We can obtain the short-run total cost curve directly from this solution:

$$STC(Q) = 16\,\frac{Q^2}{4\overline{K}} + 1\,\frac{4Q^2}{\overline{K}} + 2\overline{K} = \frac{8Q^2}{\overline{K}} + 2\overline{K}$$

The total variable and total fixed cost curves follow:

$$TVC(Q) = \frac{8Q^2}{\overline{K}}$$

$$TFC = 2\overline{K}$$

Note that, holding Q constant, total variable cost is decreasing in the quantity of capital $\overline{K}$. This is because for a given amount of output, a firm that uses more capital can typically reduce the amount of labor and raw materials it employs. Since *TVC* is the sum of expenditures on labor and materials, it follows that *TVC* should decrease in $\overline{K}$. We will see a real-world illustration of this phenomenon in Example 8.5

Similar Problem: 8.4

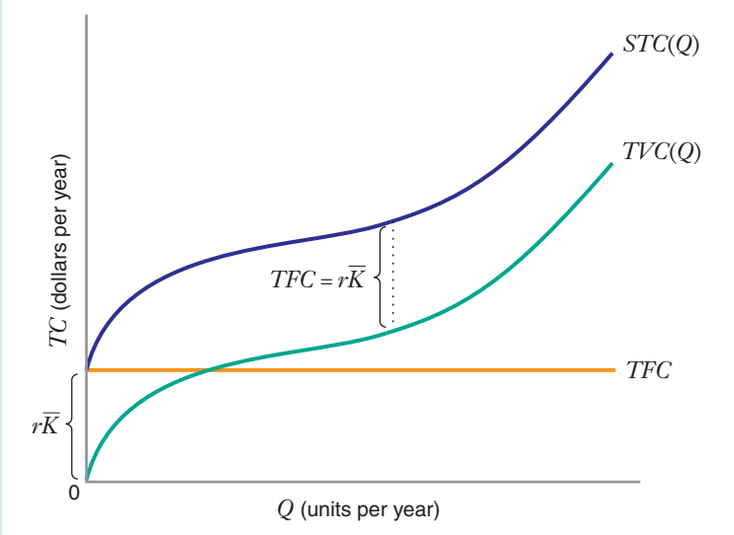

FIGURE 8.13 Short-Run Total Cost Curve
The figure shows the short-run total cost curve, *STC*(*Q*), the total variable cost curve, *TVC*(*Q*), and the total fixed cost curve, *TFC*. Total fixed cost is equal to the cost, *rK*, of the fixed capital services. Since that cost is independent of output, the total fixed cost curve is a horizontal line. At every quantity *Q*, the vertical distance between the total variable cost curve and the short-run total cost curve is equal to total fixed cost.

RELATIONSHIP BETWEEN THE LONG-RUN AND THE SHORT-RUN TOTAL COST CURVES

To develop the relationship between the long-run and short-run total cost curves, let's return to a graphical analysis of the long-run and short-run cost minimization problems for a producer of television sets. Figure 8.14 shows the relationship between the two problems when the firm uses just two inputs: labor and capital. It illustrates a point that we made in Chapter 7: When the firm is free to vary the quantity of capital in the long run, it can attain lower total costs than it can when its capital is fixed. This makes sense: the firm is more constrained when it operates in the short run, because it cannot adjust the quantity of capital freely. Specifically, suppose initially the firm wants to produce 1 million television sets, and it is free to vary both capital and labor. It would minimize total costs by operating at point A, using L_1 units of labor and K_1 units of capital. However, if the firm's desired output goes up to 2 million units, but its capital remains fixed at K_1, it would operate at point B. By contrast, long-run cost-minimization would move the firm along its expansion path to point C. Since point B is on a higher isocost line than point C, the firm incurs higher costs in the short run to produce an output of 2 million televisions than it would in the long run if it were free to vary the quantity of its capital.

Figure 8.15 shows the corresponding relationship between the long-run and short-run total cost curves. The short-run total cost curve when capital is fixed at K_1 lies everywhere above the long-run total cost curve, except at point A. This

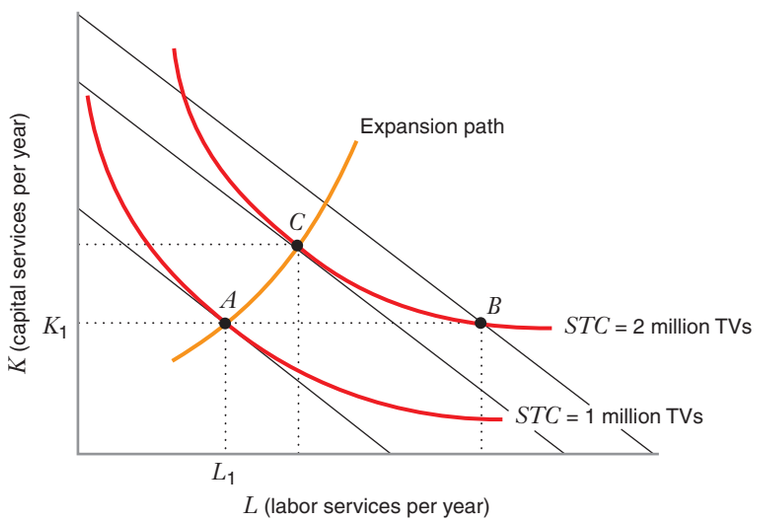

FIGURE 8.14 Why Total Costs Are Higher in the Short Run than in the Long Run Initially, the firm produces 1 million TV sets, and it minimizes long-run costs by operating at point *A*. If the firm's desired output then goes up to 2 million TV sets, but it cannot increase the quantity of capital, it must operate at point *B*. In the long run, when it can adjust the quantity of its capital, it will move from point *B* to point *C*. Since point *C* lies on a lower isocost line than *B*, the long-run total cost of producing 2 million TVs is less than the short-run total cost.

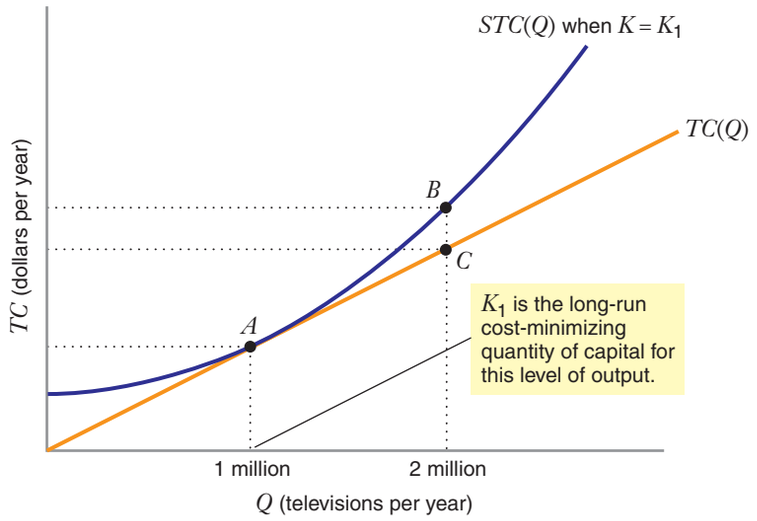

FIGURE 8.15 Relationship Between Short-Run and Long-Run Total Cost Curves
The short-run total cost curve when the quantity of capital is fixed at K_1 lies above the long-run total cost curve at every level of output, except point *A*. At this point, short-run total cost equals long-run total cost. For example, as we saw in Figure 8.14, at an output of 2 million TVs per year, the short-run total cost exceeds the long-run total cost. At point *A*, the quantity of capital K_1 is the long-run cost-minimizing quantity of capital for the output of 1 million units per year.

illustrates the point we just made: The firm cannot attain as low a level of total cost as it can in the long run when it is free to vary all its inputs. At point *A*, the short-run total cost is equal to long-run total cost. What is special about point *A*? At point *A*, the firm produces 1 million televisions per year, the quantity of output for which the fixed capital K_1 is cost minimizing in the long run. That is, at a quantity of 1 million units, the solution to the short-run cost-minimization problem when $K = K_1$ coincides with the solution to the long-run cost-minimization problem (see Figure 8.14). Therefore, at a quantity of 1 million units, short-run total cost *STC* must equal the long-run total cost *TC*.

SHORT-RUN MARGINAL AND AVERAGE COSTS

Just as we can define long-run average and long-run marginal costs, we can also define **short-run average cost** (*SAC*) and **short-run marginal cost** (*SMC*):

$$SAC(Q) = \frac{STC(Q)}{Q}$$

$$SMC(Q) = \frac{STC(Q + \Delta Q) - STC(Q)}{\Delta Q}$$

$$= \frac{\Delta STC}{\Delta Q}$$

Just as long-run marginal cost is equal to the slope of the long-run total cost curve, short-run marginal cost is equal to the slope of the short-run total cost curve. Note that in Figure 8.14 at point *A* (i.e., when output equals 1 million units per year), the slopes of the long-run total cost and short-run total cost curves are equal. It therefore follows that at this level of output, not only does $STC = TC$, but $SMC = MC$.

Because we can break short-run total cost into two pieces (total variable cost and total fixed cost), we can also break short-run average cost into two pieces: **average variable cost** (*AVC*) and **average fixed cost** (*AFC*):

$$STC = TVC + TFC, \text{ so}$$

$$SAC = AVC + AFC.$$

Put another way, average fixed cost is total fixed cost per unit of output, i.e. $AFC = TFC/Q$. Average variable cost is total variable cost per unit of output, i.e. $AVC = TVC/Q$.

Figure 8.16 illustrates typical graphs of the short-run marginal, short-run average cost, average variable cost, and average fixed cost curves. We obtain the short-run average cost curve by "vertically summing" the average variable cost curve and the average fixed cost curve.[15] The average fixed cost curve decreases everywhere and approaches the horizontal axis as *Q* becomes very large. This reflects the fact that as output increases, fixed capital costs are "spread out" over an increasingly large volume of output, driving fixed costs per unit downward. Because *AFC* becomes smaller and smaller as *Q* increases, the *AVC* and *SAC* curves get closer and closer together. The short-run marginal cost curve *SMC* intersects the short-run average cost curve and the average variable cost curve at the minimum point of each curve. This property mirrors the relationship between

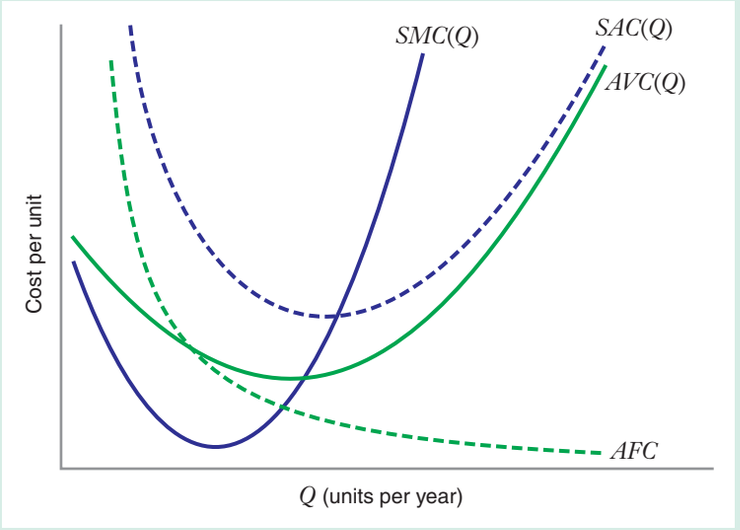

FIGURE 8.16 Short-Run Marginal and Average Cost Curves
The short-run marginal cost curve, *SMC*(*Q*), the short-run average cost curve, *SAC*(*Q*), the average variable cost curve, *AVC*(*Q*), and the average fixed cost curve, *AFC*.

[15] *Vertically summing* means that, for any *Q*, we find the height of the *SAC* curve by adding together the heights of the *AVC* and *AFC* curves at that quantity.

the long-run marginal and long-run average cost curves (and again reflects the relationship between the average and marginal measures of anything).

THE LONG-RUN AVERAGE COST CURVE AS AN ENVELOPE CURVE

Figure 8.17 illustrates the relationship between the long-run average cost curve and short-run average cost curves for a U-shaped long-run average cost curve $AC(Q)$. The figure shows different short-run average cost curves: $SAC_1(Q)$, $SAC_2(Q)$, $SAC_3(Q)$. These curves are also U-shaped. Each corresponds to a different level of fixed capital, or *plant size*, K_1, K_2, and K_3. Thinking of a television producer, such as HiSense, K_3 might either be a larger factory than K_1 or K_2, or it might entail a greater degree of automation.

A short-run average cost curve for a particular plant size lies above the long-run average cost curve except at the level of output for which that plant size is optimal. For example, a television producer, such as HiSense, that planned to produce 1 million televisions per year would minimize its production costs by building a small plant of size K_1. If it built a plant of this size and in fact produced 1 million television sets, its short-run average cost would equal the long-run average cost of \$50 per television. But if HiSense expanded its output in this small plant to, say, 2 million units, its short-run average cost would be \$110 per television,

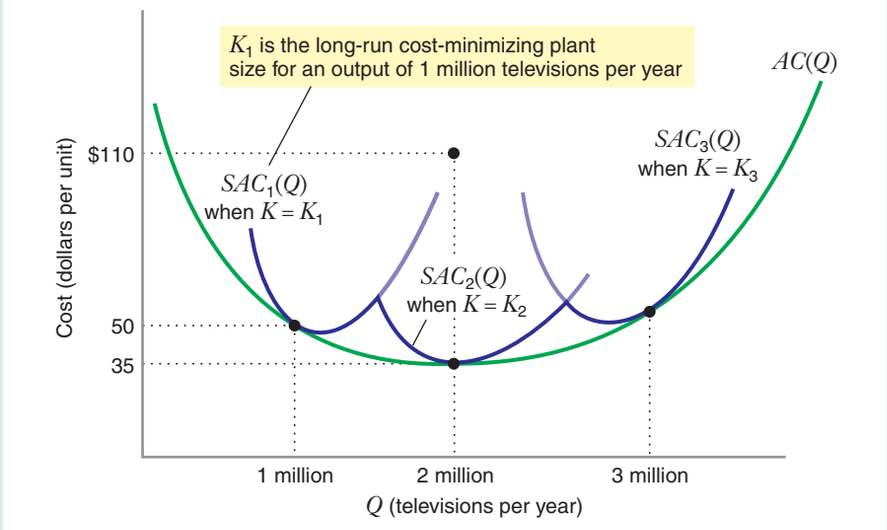

FIGURE 8.17 The Long-Run Average Cost Curve as an Envelope Curve
The figure shows three different short-run average cost curves: $SAC_1(Q)$, $SAC_2(Q)$, $SAC_3(Q)$. Each corresponds to a different level of fixed capital, or plant size. Each short-run average cost curve lies above the long-run average cost curve except at the level of output for which that plant size is cost minimizing in the long run. If we trace out the lower boundary of the three short-run average cost curves, we obtain a scalloped-shaped curve. This curve tells us the minimum attainable average cost if the firm could choose among just three plant sizes: K_1, K_2, and K_3. This scalloped curve approximates the long-run average cost curve. If we drew more short-run average cost curves and traced the lower boundary including these additional curves, we would more closely approximate the long-run average cost curve.

even though its long-run average cost at 2 million units is only $35 per television. This difference between the short-run average cost and the long-run average cost illustrates a point we made in our earlier discussion of the relationship between the short-run total cost and long-run total cost curve: you can never do better (i.e., have lower total costs) in the short run than in the long run because in the long run you can set *all* of your inputs to the levels that minimize total cost. (In practice, the high unit cost that HiSense would incur from producing a relatively large output in a small plant might reflect reductions in the marginal product of labor that arise from crowding a large work force into a small plant.) In order to attain the long-run average cost of $35 per television when producing 2 million televisions, HiSense would need to expand the size of its plant from K_1 to K_2.

If we traced the lower boundary of the three short-run average cost curves, we would obtain the dark "scalloped" curve in Figure 8.17. This curve tells us the minimum attainable average cost if the firm could choose only one of three plant sizes: K_1, K_2, and K_3. The scalloped curve approximates the long-run average cost curve. If we drew more short-run average cost curves and traced the lower boundary including these additional curves, the resulting scalloped curve would be an even better approximation to the long-run average cost curve. This argument tells us that you can think of the long-run average cost curve as the "lower envelope" of an infinite number of short-run average cost curves. The long-run average cost curve is thus sometimes referred to as the *envelope curve*.

Figure 8.18 takes Figure 8.17 one step further and shows the special relationships between the short-run average and marginal cost curves and the long-run average and marginal cost curves. At an output of 1 million units, short-run average cost equals long-run average cost. Short-run marginal cost also equals long-run marginal cost at 1 million units. These relationships reflect our earlier discussions of those between the short-run and long-run cost curves. Note, too, that since long-run average cost and long-run marginal cost are *not equal* at this particular level of output, short-run average cost and short-run marginal cost are also not equal here. (They are equal at a higher level of output.) At an output level of 3 million units, the relationships between the short-run and long-run average and marginal cost curves are analogous to those at an output of 1 million units per year.

An output of 2 million units corresponds to the point at which long-run average cost attains its minimum level—the MES. At MES, long-run marginal cost equals long-run average cost, and short-run marginal cost equals short-run average cost; that is, $AC = MC = SAC = SMC$.

LEARNING-BY-DOING EXERCISE 8.4

The Relationship Between Short-Run and Long-Run Average Cost Curves

Problem

(a) What is the long-run average cost curve for this production function?

Solution Recall from Learning-By-Doing Exercise 7.6 that the solution to the long-run cost-minimization problem is

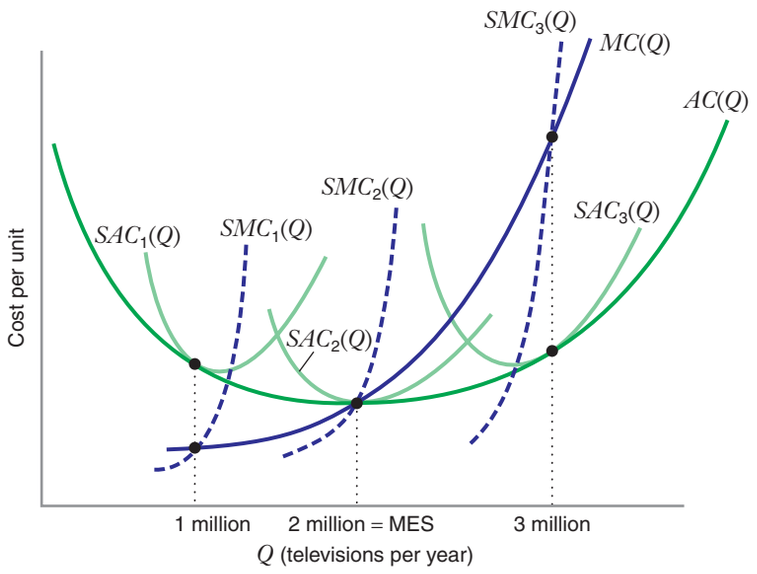

FIGURE 8.18 The Relationship Between the Long-Run Average and Marginal Cost Curves and the Short-Run Average and Marginal Cost Curves
At an output level of 1 million TVs per year (the output level for which plant size K_1 solves the long-run cost-minimization problem), short-run average cost (*SAC*) equals long-run average cost (*AC*(*Q*)). Short-run marginal cost (*SMC*) also equals long-run marginal cost (*MC*(*Q*)) at 1 million units. Since *AC*(*a*) and *MC*(*b*) are not equal at this level of output, *SAC* and *SMC* are also not equal here. At an output level of 3 million TVs per year, the relationships between the short-run and long-run average and marginal cost functions are analogous to those at 1 million units. An output of 2 million units is where *AC*(*Q*) attains its minimum level; that is, it is the *MES*. At the *MES*, *MC*(*Q*) equals *AC*(*Q*), and thus *SMC*-*SAC*.

$$L = \frac{Q}{8}.$$

$$M = 2Q.$$

$$K = 2Q.$$

The input prices are $w = 16$, $m = 1$, and $r = 2$, so the long-run total cost curve is

$$TC(Q) = 16\left(\frac{Q}{8}\right) + 1(2Q) + 2(2Q) = 8Q.$$

The long-run average cost curve is thus

$$AC(Q) = \frac{TC(Q)}{Q} = \frac{8Q}{Q} = 8.$$

Problem

(b) What is the short-run average cost curve for a fixed level of capital $\overline{K}$?

Solution We derived the short-run total cost curve for this production function in Learning-By-Doing Exercise 8.3: $STC(Q)$. Thus, the short-run average cost curve is

$$SAC(Q) = \frac{8Q}{K} + \frac{2\overline{K}}{Q}.$$

(c) Graph the long-run average cost curve and the short-run average cost curves corresponding to $\overline{K} = 10$, $\overline{K} = 20$, and $\overline{K} = 40$.

Solution The long-run average cost curve is a horizontal line, as Figure 8.19 shows. This makes sense because the production function exhibits constant returns to scale. The short-run average cost curves are U-shaped, and attain their minimum point at $Q = 5$, $Q = 10$, and $Q = 20$, respectively.

Similar Problem: 8.4

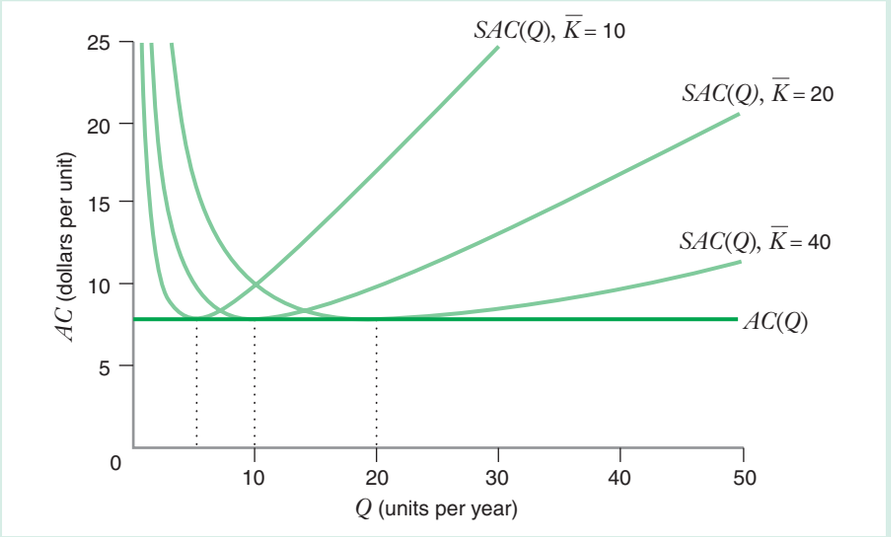

FIGURE 8.19 Long-Run and Short-Run Average Cost Curves for Learning-By-Doing Exercise 8.4
Each short-run average cost curve corresponds to a particular plant size: $K = 10$, 20, and 40. These curves are U-shaped. The long-run average cost curve is the lower envelope of the short-run average cost curves and is a horizontal line.

EXAMPLE 8.5

The Short-Run and Long-Run Cost Curves for an American Railroad Firm[16]

The 1990s were an interesting time for U.S. railroads. On the positive side, the railroad industry was profitable, and the bankruptcies that had plagued the industry in the 1960s and 1970s were over. Some railroads, such as the Norfolk Southern, had become so optimistic about the future that they had begun ambitious investments in new track. On the negative side, however, U.S. railroads had developed a generally poor reputation for service, particularly speed of delivery. On some routes, shipping freight by train in the late 1990s took longer than it did thirty years earlier. Service became so bad that in 1997 Lionel, a company that makes toy trains, began shipping by truck rather than by rail. "We feel a little guilty forsaking our big brothers," said Lionel President Gary Moreau, "But we have no choice." Part of the problem, according to industry observers, arose because the railroad industry downsized too much. During the 1980s and 1990s, U.S. railroads sold or abandoned 55,000 miles of track. According to one expert, the railroads ". . . have too much freight trying to go over too little track."

These concerns over the quality of rail service and how they relate to the amount of track a railroad employs might make you wonder how a railroad's production costs depend on these factors. For example, would a railroad's total variable costs go down as it adds track? If so, at what rate? Would a faster service increase or decrease a railroad's cost of operation?

One way to study these questions would be to estimate the short-run and long-run cost curves for a railroad. In the 1980s, Ronald Braeutigam, Andrew Daughety, and Mark Turnquist (hereafter BDT) undertook such a study.[17] With the cooperation of the management of a large American railroad firm, BDT obtained data on costs of shipment, input prices (price of fuel, price of labor service), volume of output, and speed of service for this railroad.[18] Using statistical techniques, they estimated a short-run total variable cost curve for the railroad. In the study, total variable cost is the sum of the railroad's monthly costs for labor, fuel, maintenance, car, locomotive, and supplies.

Table 8.5 shows the impact on total variable costs of a hypothetical 10 percent increase in (1) traffic volume (car-loads of freight per month); (2) the quantity of the railroad's track (in miles); (3) speed of service (miles per day of loaded cars); and (4) the prices of labor, fuel, and equipment.[19] You should think of track miles as a fixed input, analogous to capital in our previous discussion. A railroad cannot instantly vary the quantity or quality of its track to adjust to month-to-month variations in shipment volumes in the system and thus must regard track as a fixed input.

Table 8.5 contains several interesting findings. First, total variable cost increases with total output and with the prices of the railroad's inputs. This is consistent with the predictions of the theory you have been learning in this chapter and Chapter

[16]The first part of this example box draws from "A Long Haul: America's Railroads Struggle to Capture Their Former Glory," *Wall Street Journal* (December 5, 1997), p. A1 and A6.

[17]R. R. Braeutigam, A. F. Daughety, and M. A. Turnquist, "A Firm Specific Analysis of Economies of Density in the U. S. Railroad Industry," *Journal of Industrial Economics 33* (September 1984); 3–20.

[18]The identity of the firm remained anonymous to ensure the confidentiality of its data.

[19]In this study, the railroad's track mileage was adjusted to reflect changes in the quality of its track over time.

7. Second, as we discussed in Learning-By-Doing Exercise 8.3, we would expect that total variable costs would go down as the volume of the fixed input is increased. Table 8.5 shows that this is true for BDT's railroad. Holding traffic volume and speed of service fixed, an increase in track mileage (or an increase in the quality of track, holding mileage fixed) would be expected to decrease the amount the railroad spends on variable inputs, such as labor and fuel. For example, with more track (holding output and speed fixed), the railroad would reduce the congestion of trains on its mainlines and in its train yards. As a result, it would probably need fewer dispatchers to control the movement of trains. Third, Table 8.5 tells us that improvements in average speed may also reduce costs. Although this impact is not large, it does suggest that improvements in service not only can benefit the railroad's consumers, they might also benefit the railroad itself through lower variable costs. For this railroad, higher speeds might reduce use of labor (e.g., fewer train crews would be needed to haul a given amount of freight) and increase the fuel efficiency of the railroad's locomotives.

Having estimated the total variable cost function, BDT go on to estimate the long-run total and average cost curves for this railroad. They do so by finding the track mileage that, for each quantity Q, minimizes the sum of total variable costs and total fixed costs, where total fixed cost is the monthly opportunity cost to the firm's owners of a given amount of track mileage. Figure 8.20 shows the long-run average cost function estimated by BDT using this approach. It also shows two short-run average cost curves, each corresponding to a different level of track mileage. (Track mileage is stated in relation to the average track mileage observed in BDT's data.) The units of output in Figure 8.20 are expressed as a percentage of MES, and the average level of output produced by the railroad at the time of the study was about 40 percent of MES. This study thus suggests that increases in traffic volume, accompanied by cost-minimizing adjustments in track mileage, would reduce this railroad's average production costs over a wide range of output. ■

TABLE 8.5
What Affects Total Variable Costs for a Railroad?

A 10 Percent Increase In . . .	Changes Total Variable Cost By . . .
Volume of Output	+3.98%
Track Mileage	−2.71%
Speed of Service	−0.66%
Price of Fuel	+1.90%
Price of Labor	+5.25%
Price of Equipment	+2.85%

Adapted from Table 1 of R. R. Braeutigam, A. F. Daughety, and M. A. Turnquist, "A Firm Specific Analysis of Economies of Density in the U.S. Railroad Industry," *Journal of Industrial Economics, 33* (September 1984): 3–20. The percentage changes in the various factors are changes away from the average values of these factors over the period studied by BDT.

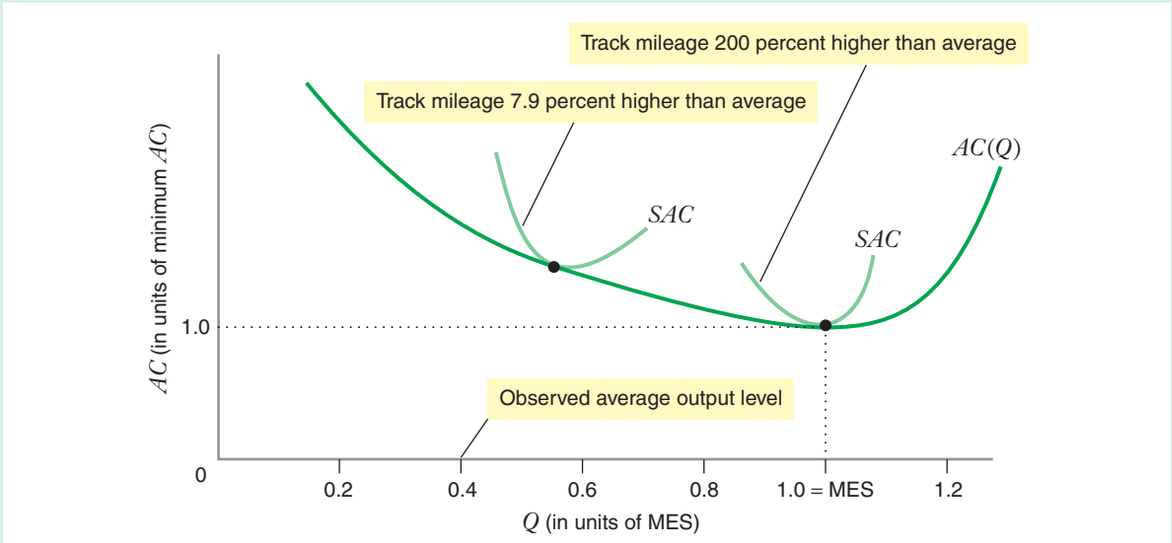

FIGURE 8.20 Long-Run and Short-Run Average Cost Curves for a Railroad
The long-run and short-run average cost curves for the railroad studied by BDT are U-shaped. The two short-run average cost curves shown correspond to a different amount of track (expressed in relation to the average amount of track observed in the data). The cost curves show that with a cost-minimizing adjustment in track size, this railroad could decrease its unit costs over a wide range of output above its current output level.

8.4 SPECIAL TOPICS IN COST

ECONOMIES OF SCOPE

This chapter has concentrated on cost curves for firms that produce just one product or service. In reality, though, many firms produce more than one product. For a firm that produces two products, total costs would depend on the quantity Q_1 of the first product the firm makes and the quantity Q_2 of the second product it makes. We will use the expression $TC(Q_1,Q_2)$ to denote how the firm's costs vary with Q_1 and Q_2. The total cost TC would be the minimized total cost of producing given quantities of the firm's two products and would come from a cost-minimization problem that is analogous to the cost-minimization problem for a single-product firm.

In some situations, efficiencies arise when a firm produces more than one product. That is, a two-product firm may be able to manufacture and market its products at a lower total cost than two single-product firms would incur when producing on a stand-alone basis. These efficiencies are called **economies of scope.**

Specifically, economies of scope exist when the total cost of producing given quantities of two goods in the same firm is less than the total cost of producing those quantities in two single-product firms. Mathematically, this definition says

$$TC(Q_1, Q_2) > TC(Q_1, 0) + TC(0, Q_2). \qquad (8.4)$$

The zeros in the expressions on the right-hand side of equation (8.4) indicate that the single-product firms produce positive amounts of one good but none of the other. These are sometimes called the **stand-alone costs** of producing goods 1 and 2.

Intuitively, the existence of economies of scope tells us that "variety" is more efficient than "specialization." We can develop an intuitive interpretation of the definition in 8.4 by rearranging terms as follows:

$$TC(Q_1, Q_2) - TC(Q_1, 0) > TC(0, Q_2) - TC(0, 0),$$

where $TC(0, 0) = 0$. That is, the total cost of producing zero quantities of both products is zero. The left-hand side of this equation is the *additional cost* of producing Q_2 units of product 2 *when the firm is already producing* Q_1 *units of product 1*. The right-hand side of this equation is the *additional cost of producing* Q_2 *when the firm does not produce* Q_1. Economies of scope exist if it is less costly for a firm to add a product to its product line given that it already produces another product. Economies of scope would exist, for example, if it is less costly for Coca-Cola to add a cherry-flavored soft drink to its product line than it would be for a new company starting from scratch.

Why would economies of scope arise? An important reason is a firm's ability to use a common input to make and sell more than one product. For example, BSkyB, the British satellite television company, can use the same satellite to broadcast a news channel, several movie channels, several sports channels, and several general entertainment channels.[20] Companies specializing in the broadcast of a single channel would each need to have a satellite orbiting the Earth. BSkyB's channels save hundreds of millions of dollars as compared to stand-alone channels by sharing a common satellite. Another example is Eurotunnel, the 31-mile tunnel that runs underneath the English Channel between Calais, France, and Dover, Great Britain. The Eurotunnel accommodates both highway and rail traffic. Two separate tunnels, one for highway traffic and one for rail traffic, would have been more expensive to construct and operate than a single tunnel that accommodates both forms of traffic.

EXAMPLE 8.6 *Nike Enters the Market for Sports Equipment*[21]

An important source of economies of scope is marketing. A company with a well-established brand name in one product line can sometimes introduce additional products at a lower cost than a stand-alone company would be able to. This is because when consumers are unsure about a product's quality they often make inferences about its quality from the product's brand name. This can give a firm with an established brand reputation an advantage over a stand-alone firm in introduc-

[20]BSkyB is a subsidiary of Rupert Murdoch's News Corporation.

[21]This example is based on "Just Doing It: Nike Plans to Swoosh Into Sports Equipment But It's a Tough Game," *Wall Street Journal* (January 6,1998), pp. A1 and A10.

ing new products. Because of its brand reputation, an established firm would not have to spend as much on advertising as the stand-alone firm to persuade consumers to try its product. This is an example of an economy of scope based on the ability of all products in a firm's product line to "share" the benefits of its established brand reputation.

A company with an extraordinary brand reputation is Nike. Nike's "swoosh," the symbol that appears on its athletic shoes and sports apparel, is one of the most recognizable marketing symbols of the modern age, and its slogan, "Just Do It," has become ingrained in American popular culture. Nike's slogan and swoosh are so recognizable that Nike can run television commercials that never mention its name and be confidant that consumers will know whose products are being advertised.

In the late 1990s, Nike turned its attention to the sports equipment market, introducing products such as hockey sticks and golf balls. Nike's goal was to become the dominant firm in the $40 billion per year sports equipment market by 2005. This is a bold ambition. The sports equipment market is highly fragmented, and no single company has ever dominated the entire range of product categories that Nike intends to enter. In addition, while no one can deny Nike's past success in the athletic shoe and sports apparel markets, producing a high-quality hockey stick or an innovative golf ball has little in common with making sneakers or jogging clothes. It therefore seems unlikely that Nike could attain economies of scope in manufacturing or product design.

Nike hopes to achieve economies of scope in marketing. These economies of scope would be based on its incredibly strong brand reputation, its close ties to sports equipment retailers, and its special relationships with professional athletes such as Tiger Woods and Ken Griffey, Jr. Nike's plan is to develop sports equipment that it can claim is innovative and then use its established brand reputation and its ties with the retail trade to convince consumers that its products are technically superior to existing products. If this plan works, Nike will be able to introduce its new products at far lower costs than a stand-alone company would incur to introduce otherwise identical products.

It will be interesting to see whether Nike succeeds. Economies of scope in marketing can be powerful, but they also have their limits. A strong brand reputation can induce consumers to try a product once, but if it does not perform as expected or if its quality is inferior, it may be difficult to penetrate the market or get repeat business. Nike's preliminary forays into the sports equipment market illustrate this risk. In July 1997, Nike "rolled out" a new line of roller skates at the annual sports equipment trade show in Chicago. But when a group of skaters equipped with Nike skates rolled into the parking lot, the wheels on the skates began to disintegrate! Quality problems have also arisen with a line of ice skates that Nike introduced several years ago. Jeremy Roenick, a star with the Phoenix Coyote's NHL hockey team, turned down a six-figure endorsement deal with Nike because he felt the skates were poorly designed and did not fit properly. Rumor has it that other hockey players who do have equipment deals with Nike use the products of competitors. According to one NHL equipment manager, "They're still wearing the stuff they've been wearing for years. They just slap the swoosh on it." ■

ECONOMIES OF EXPERIENCE: THE EXPERIENCE CURVE

Learning-by-Doing and the Experience Curve

Economies of scale refer to the cost advantages that flow from producing a larger output at a given point in time. **Economies of experience** refer to cost advantages that result from accumulated experience, or as its sometimes called, *learning-by-doing*. This is the reason we gave that title to the exercises in this book—they are designed to help you *learn* microeconomics *by doing* microeconomics problems.

Economies of experience arise for several reasons. Workers often improve their performance of specific tasks by performing them over and over again. Engineers often perfect product designs as they accumulate know-how about the manufacturing process. Firms often become more adept at handling and processing materials as they deepen their production experience. The benefits of learning are usually greater labor productivity (more output per unit of labor input), fewer defects, and higher material yields (more output per unit of raw material input).

Economies of experience are described by the **experience curve,** a relationship between average variable cost and cumulative production volume.[22] A firm's cumulative production volume at any given time is the total amount of output that it has produced over the history of the product until that time. For example, if Boeing's output of 777 jet aircraft was 30 in 1993, 45 in 1994, 50 in 1995, 70 in 1996, and 60 in 1997, its cumulative output as of the beginning of 1998 would be 30 + 45 + 50 + 70 + 60, or 255 aircraft. A typical relationship between average variable cost and cumulative output is

$$AVC(N) = AN^B,$$

where AVC is the average variable cost of production and N denotes cumulative production volume. In this formulation, A and B are constants, where $A > 0$ and B is a negative number between -1 and 0. The constant A represents the average variable cost of the first unit produced, and B represents the **experience elasticity:** the percentage change in average variable cost for every 1 percent increase in cumulative volume.

The magnitude of cost reductions that are achieved through experience is often expressed in terms of a concept known as the **slope of the experience curve.**[23] The slope of the experience curve tells us how much average variable costs go down as a percentage of an initial level when cumulative output doubles.[24] For example, if doubling a firm's cumulative output of semiconductors results in average variable cost falling from $10 per megabyte to $8.50 per megabyte, we would say that the slope of the experience curve for semiconductors is 85 percent, since average variable costs fell to 85 percent of their initial level. In terms of an equation,

$$slope = \frac{AVC(2N)}{AVC(N)}.$$

[22]The experience curve is also known as the learning curve.

[23]The slope of the experience curve is also known as the progress ratio.

[24]Note that the term slope as used here is *not* the usual notion of the slope of a straight line.

The slope and the experience elasticity are systematically related. If the experience elasticity is equal to B, the slope turns out to equal 2^B. Figure 8.21 shows experience curves with three different slopes: 90 percent, 80 percent, and 70 percent. The smaller the slope, the "steeper" the experience curve, (i.e., the more rapidly average variable costs fall as the firm accumulates experience). Note, though, that all three curves eventually flatten out. For example, beyond a volume of $N = 40$, increments in cumulative experience have a small impact on average variable costs, no matter what the slope of the experience curve is. At this point, most of the economies of experience are exhausted.

Experience curve slopes have been estimated for many different products. The median slope appears to be about 80 percent, implying that for the typical firm, each doubling of cumulative output reduces average variable costs to 80 percent of what they were before. Slopes vary from firm to firm and industry to industry, however, so that the slope enjoyed by any one firm for any given production process generally falls between 70 and 90 percent and may be as low as 60 percent or as high as 100 percent (i.e., no economies of experience).

Economies of Experience versus Economies of Scale

Economies of experience differ from economies of scale. Economies of scale refer to the ability to perform activities at a lower unit cost when those activities are performed on a larger scale. Economies of experience refer to reductions in unit costs due to accumulating experience. Economies of scale may be substantial even when learning economies are minimal. This is likely to be the case in mature, capital-intensive production processes, such as aluminum can manufacturing. Likewise, economies of experience may be substantial even when economies of scale are minimal, as in complex labor-intensive activities such as the production of handmade watches.

Firms that do not correctly distinguish between economies of scale and experience might draw incorrect inferences about the benefits of size in a market. For example, if a firm has low average costs because of economies of scale,

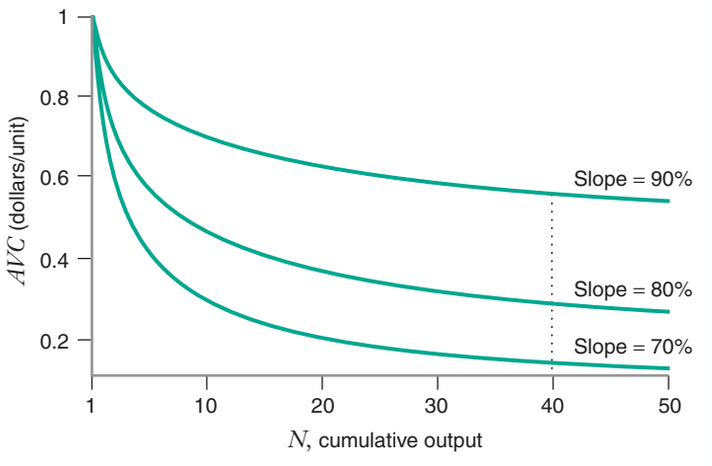

FIGURE 8.21 Experience Curves With Different Slopes
The figure shows three experience curves, each with a different slope. The smaller the slope, the "steeper" the experience curve, and the more rapidly average variable costs fall as cumulative experience goes up. No matter what the slope, though, once cumulative experience becomes sufficiently large (e.g., $N = 40$), additional increments to experience do not lower average variable costs by much.

reductions in the current volume of production will increase unit costs. If the low average costs are the result of cumulative experience, the firm may be able to cut back current production volumes without necessarily raising its average costs.

EXAMPLE 8.7 *The Experience Curve in the Production of EPROM Chips*[25]

An interesting example of economies of experience occurs in the production of semiconductors, the memory chips that are used in personal computers, cellular telephones, and electronic games. It is widely believed that the "yield" of semiconductor chips—the ratio of usable chips total chips on a silicon wafer—goes up as a firm gains production experience.[26] Silicon is an expensive raw material, and the cost of a chip is primarily determined by how much silicon it uses. The rate at which yields go up with experience is thus important for a semiconductor manufacturer to know.

Harald Gruber estimated the experience curve for a particular type of semiconductor: erasable programmable read only memory (EPROM) chips. EPROM chips are used to store program code for cellular phones, pagers, modems, video games, printers, and hard disk drives. An EPROM chip differs from the more common DRAM in that it is *nonvolatile*, which means that unlike a DRAM chip it retains its stored data when the power is turned off. And in contrast to DRAM chips, which are produced by large semiconductor firms such as Samsung and NEC, EPROM chips are generally produced by smaller firms, such as the Taiwanese firm Macronix.

Gruber recognized that other factors, such as economies of scale and memory capacity, could influence the average cost of producing an EPROM chip. After controlling for these factors, Gruber found evidence of economies of experience in the production of EPROM chips. His estimate of the slope of the EPROM experience curve was 78 percent. Thus, by doubling its cumulative volume of chips, an EPROM producer would expect its average variable costs to fall to 78 percent of their initial level.

This is an interesting finding. The market for EPROM chips is smaller than markets for other semiconductors, such as DRAMs, and as mentioned, most firms operate on a small scale. Moreover, new generations of EPROM chips are introduced frequently, typically about once every 18 months. By contrast new generations of DRAM chips were introduced about every 3 years during the 1980s and 1990s. This suggests that it is unlikely that an EPROM manufacturer will operate on the "flat" portion of the experience curve for long. By the time a firm starts to "move down" the experience curve, a new generation of chip will have come along. This, then, implies that a firm, such as Macronix, that can achieve a head start in bringing a new generation of EPROM chips to market, may achieve a significant cost advantage over slower competitors. ■

[25]This example draws from H. Gruber, "The Learning Curve in the Production of Semiconductor Memory Chips," *Applied Economics, 24* (August 1992): 885–894.

[26]A *wafer* is a slice of polycrystalline silicon. A chip producer will etch hundreds of circuits onto a single wafer.

8.5 ESTIMATING COST FUNCTIONS*

Suppose you wanted to estimate how the total costs for a television producer, such as HiSense, varied with the quantity of its output or the magnitude of its input prices. To do this, you might want to estimate what economists call a **total cost function.** A total cost function is a mathematical relationship that shows how total costs vary with the factors that influence total costs. These factors are sometimes called **cost drivers.** We've spent much of this chapter analyzing two key cost drivers: input prices and scale (volume of output). Our discussion in the previous section suggests two other cost drivers that could also influence total costs: scope (variety of other goods produced by the firm) and cumulative experience.

How would you estimate a cost function? You would first need to gather data. When estimating cost functions, many economists use data from a cross-section of firms or plants at a particular point in time. A cross-section of television producers would consist of a sample of manufacturers or manufacturing facilities in a particular year, such as 2001. For each observation in your cross-section, you would need information about total costs and cost drivers. The set of *cost drivers* that you include in your analysis is usually specific to what you are studying. In television manufacturing, scale, cumulative experience, labor wages, materials prices, and costs of capital would probably be important drivers for explaining the behavior of average costs in the long run.

Having gathered data on total costs and cost drivers, you would then use statistical techniques to construct an estimated total cost function. The most common technique used by economists is multiple regression. The basic idea behind this technique is to find a function that best fits our available data.

CONSTANT ELASTICITY COST FUNCTION

An important issue when you use multiple regression to estimate a cost function is the functional form that relates the dependent variable of interest—in this case, total cost—to the independent variables of interest, such as output and input prices. One common functional form is **constant elasticity cost function.** A constant elasticity cost function specifies a multiplicative relationship between total cost, output, and input prices.

For a production process that involves two inputs, capital and labor, the constant elasticity long-run total cost function is

$$TC = aQ^b w^a r^d,$$

where a, b, c, and d are constants. It is common to convert this into a relationship that is linear in the logs:

$$\log TC = \log a + b \log Q + c \log w + d \log r,$$

and in this form, the constants a, b, c, and d can be estimated using multiple regression.

A useful feature of the constant elasticity specification is that the constant b is the output elasticity of total cost, discussed earlier. Analogously, the constants c and d are the elasticities of long-run total cost with respect to the prices of labor and capital. These elasticities must be positive since, as we saw earlier, an increase in an input price will increase long-run total cost. We also learned

earlier that a given percentage increase in w and r would have to increase long-run total cost by the same percentage amount. This implies that the constants c and d must add up to 1 (i.e., $c + d = 1$). Thus, for the estimated long-run total cost function to be consistent with long-run cost minimization, this restriction would have to hold. This restriction can be readily incorporated into the multiple regression analysis.

TRANSLOG COST FUNCTION

The constant elasticity cost function does not allow for the possibility of average costs that first decrease in Q and then increase in Q (i.e., economies of scale, followed by diseconomies of scale). A cost function that allows for this possibility is the **translog cost function.** A translog cost function postulates a quadratic relationship between the log of total cost and the logs of input prices and output. The equation of the translog cost function is

$$\begin{aligned}\log TC = {} & b_0 + b_1 \log Q + b_2 \log w + b_3 \log r + b_4 (\log Q)^2 + \\ & + b_5(\log w)^2 + b_6 (\log r)^2 + b_7 (\log w)(\log r) \\ & + b_8 (\log w)(\log Q) + b_9 (\log r)(\log Q).\end{aligned} \qquad \textbf{(8.5)}$$

This formidable-looking expression turns out to have a lot of useful properties. For one thing, it is often a good approximation of the cost functions that come from just about *any* production function. Thus, if (as is often the case) we don't know the exact functional form of the production function, the translog would be a good choice for the functional form of the cost function. In addition, the average cost function for the translog total cost function can be U-shaped. Thus, it allows for both economies of scale and diseconomies of scale. Note, too, that if $b_5 = b_6 = b_7 = b_8 = b_9 = 0$, the translog cost function reduces to the constant elasticity cost function. Thus, the constant elasticity cost function is a special case of the translog cost function. Finally, the restrictions on the constants that make a percentage increase in all input prices lead to the same percentage increase in long-run total cost (so that the cost function is consistent with long-run cost minimization) are not difficult to state. For the cost function in (8.5) they are as follows:

$$\begin{aligned} b_2 + b_3 &= 1 \\ b_5 + b_6 + b_7 &= 0 \\ b_8 + b_9 &= 0 \end{aligned}$$

CHAPTER SUMMARY

- The long-run total cost curve shows how the minimized level of total cost varies with the quantity of output. **LBD 8.1**
- An increase in factor prices rotates the long-run total cost curve upward through the point $Q = 0$.
- Long-run average cost is the firm's cost per unit of output. It equals total cost divided by output. **LBD 8.2**
- Long-run marginal cost curve is the rate of change of long-run total cost with respect to output. **LBD 8.2**
- Economies of scale describe a situation in which long-run average cost decreases in output. Economies of scale arise because of the physical properties of processing units, specialization of labor, and indivisibilities.

- Diseconomies of scale describe a situation in which long-run average cost increases in output. A key source of diseconomies of scale are managerial diseconomies.
- The minimum efficient scale (MES) is the smallest quantity at which the long-run average cost curve attains its minimum.
- The output elasticity of total cost is the percentage change in total cost per 1 percent change in output.
- The short-run total cost curve tells us the minimized total cost as a function of output, input prices, and the level of the fixed input(s). **LBD 8.3**
- Short-run total cost is the sum of two components: total variable cost and total fixed cost.
- Corresponding to the short-run total cost curve are the short-run average cost and short-run marginal cost curves. Short-run average cost is the sum of average variable cost and average fixed cost.
- The long-run average cost curve is the lower envelope of short-run average cost curves. **LBD 8.4**
- Economies of scope exist when it is less costly to produce given quantities of two products with one firm than it is with two firms, each specializing in the production of a single product.
- Economies of experience exist when average variable cost decreases with cumulative production volume. The experience curve tells us how average variable costs are affected by changes in cumulative production volume.
- Cost drivers are factors such as output or the prices of inputs that influence the level of costs.
- Two common functional forms that are used for real-world estimation of cost functions are the constant elasticity cost function, and the translog cost function.

REVIEW QUESTIONS

1. What is the relationship between the solution to the firm's long-run cost minimization problem and the long-run total cost curve?

2. Explain why an increase in the price of an input must typically cause an increase in the long-run total cost of producing any particular level of output.

3. If the price of labor increases by 20 percent, but all other input prices remain the same, would the long-run total cost at a particular output level go up by more than 20 percent, less than 20 percent, or exactly 20 percent? If the prices of all inputs went up by 20 percent, would long-run total cost go up by more than 20 percent, less than 20 percent, or exactly 20 percent?

4. How would an increase in the price of labor shift the long-run *average* cost curve?

5. a) If the *average* cost curve is increasing, must the average cost curve always lie above the marginal cost curve? Why or why not?
b) If the *marginal* cost curve is increasing, must the average cost curve always lie above the marginal cost curve? Why or why not?

6. Sketch the long-run marginal cost curve for the "flat-bottomed" long-run average cost curve shown in Figure 8.11.

7. Could the output elasticity of total cost ever be negative?

8. Explain why the short-run marginal cost curve must intersect the average variable cost curve at the minimum point of the average variable cost curve.

9. Suppose the graph of the average variable cost curve is flat. What shape would the short-run marginal cost curve be? What shape would the short-run average cost curve be?

10. Suppose that the minimum level of short-run average cost was the same for every possible plant size. What would that tell you about the shapes of the long-run average and long-run marginal cost curves?

11. What is the difference between economies of scope and economies of scale? Is it possible for a two-product firm to enjoy economies of scope but not economies of scale? Is it possible for a firm to have economies of scale but not economies of scope?

12. What is an experience curve? What is the difference between economies of experience and economies of scale?

PROBLEMS

1. A firm produces a product with labor and capital, and its production function is described by

$$Q = LK.$$

The marginal products associated with this production function are

$$MP_L = K.$$

$$MP_K = L.$$

Suppose that the price of labor equals 2, and the price of capital equals 1. Derive the equations for the long-run total cost curve and the long-run average cost curve.

2. A firm's long-run total cost curve is

$$TC(Q) = 1000Q + 30Q^2 - Q^3.$$

Derive the expression for the corresponding long-run average cost curve and then sketch it. At what quantity is minimum efficient scale?

3. Consider a production function of two inputs: labor and capital, given by

$$Q = [L^{\frac{1}{2}} + K^{\frac{1}{2}}]^2.$$

The marginal products associated with this production function are as follows:

$$MP_L = [L^{\frac{1}{2}} + K^{\frac{1}{2}}]\, L^{-\frac{1}{2}}.$$

$$MP_K = [L^{\frac{1}{2}} + K^{\frac{1}{2}}]\, K^{-\frac{1}{2}}.$$

Let $w = 2$ and $r = 1$.

a) Suppose the firm is required to produce Q units of output. Show how the cost-minimizing quantity of labor depends on the quantity Q. Show how the cost-minimizing quantity of capital depends on the quantity Q.
b) Find the equation of the firm's long-run total cost curve.
c) Find the equation of the firm's long-run average cost curve.
d) Find the solution to the firm's short-run cost-minimization problem when capital is fixed at a quantity of 10 units (i.e., $\overline{K} = 9$).
e) Find the short-run total cost curve, and graph it along with the long-run total cost curve.
f) Find the associated short-run average cost curve.

4. Consider a production function of three inputs, labor, capital, and materials given by

$$Q = LKM$$

The marginal products associated with this production function are as follows:

$$MP_L = KM$$

$$MP_K = LM$$

$$MP_M = LK$$

Let $w = 5$, $r = 1$, and $m = 2$, where m is the price per unit of materials.

a) Suppose that the firm is required to produce Q units of output. Show how the cost-minimizing quantity of labor depends on the quantity Q. Show how the cost-minimizing quantity of capital depends on the quantity Q. Show how the cost-minimizing quantity of materials depends on the quantity Q.
b) Find the equation of the firm's long-run total cost curve.
c) Find the equation of the firm's long-run average cost curve.
d) Suppose that the firm is required to produce Q units of output, but that its capital is fixed at a quantity of 50 units (i.e., $\overline{K} = 50$). Show how the cost-minimizing quantity of labor depends on the quantity Q. Show how the cost-minimizing quantity of materials depends on the quantity Q.
e) Find the equation of the short-run total cost curve when capital is fixed at a quantity of 50 units (i.e., $\overline{K} = 50$) and graph it along with the long-run total cost curve.
f) Find the equation of the associated short-run average cost curve.

5. A short-run total cost curve is given by the equation

$$STC(Q) = 1000 + 50Q^2.$$

Derive expressions for and then sketch the corresponding short-run average cost, average variable cost, and average fixed cost curve.

6. A producer of hard disk drives has a short-run total cost curve given by

$$STC(Q) = \overline{K} + \frac{Q^2}{\overline{K}}.$$

Within the same set of axes, sketch a graph of the short-run average cost curves for three different plant sizes: $\overline{K} = 10$, $\overline{K} = 20$, and $\overline{K} = 30$. Based on this graph, what is shape of the long-run average cost curve?

7. Figure 8.17 shows that the short-run marginal cost curve may lie above the long-run marginal cost curve. Yet, in the long run, the quantities of all inputs are variable, whereas in the short run, the quantities of just some of the inputs are variable. Given that, why isn't short-run marginal cost less than long-run marginal cost for all output levels?

8. Suppose that the total cost of providing satellite television services is as follows

$$TC(Q_1, Q_2) = \begin{cases} 0 & \text{if } Q_1 = 0 \text{ and } Q_2 = 0. \\ 1000 + 2Q_1 + 3Q_2 & \text{otherwise,} \end{cases}$$

where Q_1 and Q_2 are the number of households that subscribe to a sports and movie channel, respectively. Does the provision of satellite television services exhibit economies of scope?

9. A researcher has claimed to have estimated a long-run total cost function for the production of automobiles. His estimate is that

$$TC(Q, w, r) = 100w^{-\frac{1}{2}}r^{\frac{1}{2}}Q^3,$$

where w and r are the prices of labor and capital. Is this a valid cost function—that is, it consistent with long-run cost minimization by the firm? Why or why not?

APPENDIX: Shephard's Lemma and Duality

WHAT IS SHEPHARD'S LEMMA?

Let's compare our calculations in Learning-By-Doing Exercise 7.4 in Chapter 7 and Learning-By-Doing Exercise 8.1 in this chapter. Both pertain to the production function $Q = 50K^{\frac{1}{2}}L^{\frac{1}{2}}$. Our input demand functions were

$$K^*(Q, w, r) = \frac{Q}{50}\left(\frac{w}{r}\right)^{\frac{1}{2}}.$$

$$L^*(Q, w, r) = \frac{Q}{50}\left(\frac{r}{w}\right)^{\frac{1}{2}}.$$

Our long-run total cost function was

$$TC(Q, w, r) = \frac{w^{\frac{1}{2}}r^{\frac{1}{2}}}{25}Q.$$

Let's see how the long-run total cost function varies with respect to the price of labor w, holding Q and r fixed.

$$\frac{\partial TC(Q, w, r)}{\partial w} = \frac{Q}{50}\left(\frac{r}{w}\right)^{\frac{1}{2}} = L^*(Q, w, r). \qquad \textbf{(A8.1)}$$

The rate of change of long-run total cost with respect to the price of labor is equal to the labor demand function. Similarly,

$$\frac{\partial TC(Q, w, r)}{\partial r} = \frac{Q}{50}\left(\frac{r}{w}\right)^{\frac{1}{2}} = K^*(Q, w, r). \qquad \textbf{(A8.2)}$$

The rate of change of long-run total cost with respect to the price of capital is equal to the capital demand function.

The relationships summarized in equations (A8.1) and (A8.2) are no coincidence. They reflect a general relationship between the long-run total cost function

and the input demand functions. This relationship is known as **Shephard's Lemma.** Shephard's Lemma states that the *rate of change of long-run total cost function with respect to an input price is equal to the corresponding input demand function.*[27] Mathematically,

$$\frac{\partial TC(Q, w, r)}{\partial w} = L^*(Q, w, r).$$

$$\frac{\partial TC(Q, w, r)}{\partial r} = K^*(Q, w, r).$$

Shephard's Lemma makes intuitive sense: if a firm experienced an increase in its wage rate by $1 per hour, then its total costs should go up (approximately) by the $1 increase in wages multiplied by the amount of labor it is currently using; i.e., the rate of increase in total costs should be approximately equal to its labor demand function. We say "approximately" because if the firm minimizes its total costs, the increase in w should cause the firm to decrease the quantity of labor and increase the quantity of capital it uses. Shephard's Lemma tells us that for small enough changes in w (i.e., Δw sufficiently close to 0), we can use the firm's current usage of labor as a good approximation for how much a firm's costs will rise.

DUALITY

What is the significance of Shephard's Lemma? It provides a key link between the production function and the cost function, a link that in the Appendix to Chapter 7 we called duality. Duality works like this:

- Shephard's Lemma tells us that if we know the total cost function, we can derive the input demand functions.
- In turn, as we saw in the Appendix to Chapter 7, if we know the input demand functions, we can "back out" the production function.

Thus, if we know the total cost function, we can always "back out" the production function from which it must have been derived. In this sense, the cost function is *dual* (i.e., linked) to the production function. For any production function, there is a unique total cost function that can be derived from it via the cost minimization problem. And if we know that total cost function, we can recover the production function that is "dual" to it.

This is a valuable insight. Estimating a firm's production function by statistical methods is often difficult. For one thing, among the many choices of "specific" functional forms for a production function, how would you know which one is most appropriate for a particular industry or firm? In addition, data on input prices and total costs are often more readily available than data on the quantities of inputs. An example of research that took advantage of Shephard's Lemma

[27]Shephard's Lemma also applies to the relationship between short-run total cost functions and the short-run input demand functions. For that reason, we will generally not specify whether we are in the short run or long run in the remainder of this section. However, to maintain a consistent notation, we will use the "long-run" notation used in this chapter and Chapter 7.

is the studies of economies of scale in electricity power generation discussed in Example 6.3. In these studies, the researchers estimated cost functions using statistical methods. They then applied Shephard's Lemma and the logic of duality to infer the nature of returns to scale in the production function.

HOW DO TOTAL, AVERAGE, AND MARGINAL COST VARY WITH INPUT PRICES

We can use Shepard's Lemma to determine how the total, average, and marginal cost functions vary with input prices. Total and average cost are easy. For any $Q > 0$, Shephard's Lemma tells us that total cost $TC(Q,w,r)$ must go up as an input price goes up, provided that the firm uses a positive amount of the input. Using the price w of labor as an example, this is because:

$$\frac{\partial TC(Q, w, r)}{\partial w} = L^*(Q, w, r) > 0.$$

And because average cost is total cost divided by quantity, it follows that

$$\frac{\partial AC(Q, w, r)}{\partial w} = \frac{L^*(Q, w, r)}{Q} > 0.$$

Thus, average cost must also increase as an input price goes up.

The impact of an input price on marginal cost is trickier. Recall that marginal cost is the rate of change of total cost with respect to Q, or:

$$MC(Q, w, r) = \frac{\partial TC(Q, w, r)}{\partial Q}.$$

Thus, we express the rate of change of marginal cost with respect to an input price, such as w, this way:

$$\begin{aligned}\frac{\partial MC(Q, w, r)}{\partial w} &= \frac{\partial^2 TC(Q, w, r)}{\partial w \partial Q} \\ &= \frac{\partial\left(\frac{\partial TC(Q,w,r)}{\partial w}\right)}{\partial Q} \\ &= \frac{\partial L^*(Q, w, r)}{\partial Q}.\end{aligned}$$

The last line in the above expression is a consequence of Shephard's Lemma since

$$\frac{\partial TC(Q, w, r)}{\partial w} = L^*(Q, w, r),$$

Thus, Shephard's Lemma implies that the *rate of change of marginal cost with respect to the price of an input (e.g., labor) is equal to the rate of change of the demand for that input (e.g., labor) with respect to output.* It then follows that

- An increase in the price of a *normal input* (input demand increases in output Q) will *increase* marginal cost.[29]
- An increase in the price of an *inferior input* (input demand decreases in output Q) will *decrease* marginal cost.

We can now summarize what Shephard's Lemma tells us about the relationship between input prices and the cost functions:

- An increase in an input price will increase total cost TC as long as quantity Q is positive and the firm uses a positive quantity of the input.
- An increase in an input price will increase average cost AC as long as quantity Q is positive and the firm uses a positive quantity of the input.
- An increase in an input price will increase marginal cost MC if the input is normal input, and it will decrease marginal cost if the input is inferior.

A decrease in the price of an input will affect total, average, and marginal cost in an analogous manner.

PROOF OF SHEPHARD'S LEMMA

For a fixed Q, let L_0 and K_0 be the cost minimizing input combination for any arbitrary combination of input prices w_0, r_0,

$$L_0 = L^*(Q, w_0, r_0).$$
$$K_0 = K^*(Q, w_0, r_0).$$

Now define a function of w and r, $g(w, r)$ equal to

$$g(w, r) = TC(Q, w, r) - wL_0 - rK_0.$$

What is special about this function? Well, we know that since L_0, K_0 is the cost minimizing input combination when $w = w_0$ and $r = r_0$, it must be the case that

$$g(w_0, r_0) = 0. \qquad \textbf{(A8.3)}$$

Moreover, since L_0, K_0 is a feasible (but possibly non-optimal) input combination to produce output Q at other input prices w, r besides w_0, r_0, it must be the case that

$$g(w, r) \leq 0 \text{ for } (w, r) \neq (w_0, r_0). \qquad \textbf{(A8.4)}$$

[28]See Chapter 7 to review the concepts of normal and inferior inputs.

Conditions (A8.3) and (A8.4) imply that the function $g(w, r)$ attains its maximum when $w = w_0$ and $r = r_0$. Hence, at these points, its partial derivatives with respect to w and r must be zero:

$$\frac{\partial g(w_0, r_0)}{\partial w} = 0 \longrightarrow \frac{\partial TC(Q, w_0, r_0)}{\partial w} = L_0. \quad \textbf{(A8.5)}$$

$$\frac{\partial g(w_0, r_0)}{\partial r} = 0 \longrightarrow \frac{\partial TC(Q, w_0, r_0)}{\partial r} = K_0. \quad \textbf{(A8.6)}$$

But since $L_0 = L^*(Q, w_0, r_0)$ and $K_0 = K^*(Q, w_0, r_0)$, (A8.5) and (A8.6) imply

$$\frac{\partial TC(Q, w_0, r_0)}{\partial w} = L^*(Q, w_0, r_0). \quad \textbf{(A8.7)}$$

$$\frac{\partial TC(Q, w_0, r_0)}{\partial r} = K^*(Q, w_0, r_0). \quad \textbf{(A8.8)}$$

Since w_0, r_0 is an arbitrary combination of input prices, conditions (A8.7) and (A8.8) hold for any pair of input prices, and this is exactly what we wanted to show to prove Shephard's Lemma.