CHAPTER

6 Production Theory and Analysis

PREVIEW

A firm is an entity that combines and processes resources in order to produce output that will directly or indirectly satisfy consumer demand. Firms range in size from the person who bakes apple pies at home for sale to neighbors to the largest multinational conglomerate. However, when reduced to the basics, all firms do the same thing—they employ resources to produce output that will be sold in the market. The goal of the firm is to maximize the profit earned from this activity.

The general production problem facing the firm is to determine how much output to produce and how much labor and capital to employ to produce that output most efficiently. Engineering information in the form of a production function and economic information on prices of outputs and inputs must be combined in order to answer those questions.

In this chapter, a framework for understanding the economics of production is presented and a set of conditions for efficient production developed. In the first section, the concept of a production function is outlined. In following sections, techniques for determining cost-efficient production and input rates are determined. This is done for one variable input and then for two variable inputs. Next, the concepts of returns to scale and economies of scope are discussed. Finally, techniques for statistical estimation of production functions are developed.

THE PRODUCTION FUNCTION

For simplicity, assume that all inputs or factors of production can be grouped into two broad categories, labor (L) and capital (K). The general equation for the production function is

$$Q = f(K, L) \tag{6-1}$$

This function defines the maximum rate of output (Q) per unit of time obtainable from a given rate of capital and labor input. Output may be in physical units such as automobiles or microcomputers, or it may be intangible, as in the case of medical care, transportation, or education.

The production function is really an engineering concept that is devoid of economic content. That is, it simply relates output and input rates. The production function does not yield information on the least-cost capital–labor combination for producing a given level of output, nor does it reveal the output rate that would yield maximum profit. The function only shows the maximum output obtainable from any and all input combinations. Prices of the inputs and the price of output must be used with the production function to determine which of the many possible input combinations is best, given the firm's objective.

The definition of the production function as defining maximum output rates is important. Obviously, firms can fail to organize or manage resources efficiently and produce less than the maximum output for given input rates. However, in a competitive environment, such firms are not likely to survive because competitors using efficient production techniques will be able to produce at lower cost, sell at lower prices, and ultimately drive inefficient producers out of the market. Thus, only firms using the best production methods (i.e., maximizing production from any input combination) are considered.

Economists use a variety of functional forms to describe production. The multiplicative form, generally referred to as a Cobb–Douglas production function,

$$Q = AK^{\alpha}L^{\beta} \tag{6-2}$$

is widely used in economics because it has properties representative of many production processes. It will be used as the basis for many of the examples found in this chapter.

Consider a Cobb–Douglas production function with parameters $A = 100$, $\alpha = 0.5$, and $\beta = 0.5$. That is,

$$Q = 100K^{0.5}L^{0.5} \tag{6-3}$$

A production table shows the maximum rate of output associated with each of a number of input combinations. For example, given the production function (6-3), if two units of labor and four units of capital are used, maximum production is 283 units of output. If $K = 8$ and $L = 2$ the output rate will be 400. Table 6.1 shows production rates for various input-rate combinations applied to the production function (6-3).

Three important relationships are shown by the data in this production table. First, the table indicates that there are a variety of ways to produce a particular rate of output. For example, 245 units of output can be produced with any of the following input combinations:

Combination	*K*	*L*
a	6	1
b	3	2
c	2	3
d	1	6

This implies that there is *substitutability* between the factors of production. The firm can use a capital-intensive production process characterized by combination *a*, a labor-intensive process such as *d*, or a process that uses a resource combination somewhere

Rate of Capital Input (K)								
8	283	400	490	565	632	693	748	800
7	265	374	458	529	592	648	700	748
6	245	346	424	490	548	600	648	693
5	224	316	387	447	500	548	592	632
4	200	283	346	400	447	490	529	565
3	173	245	300	346	387	424	458	490
2	141	200	245	283	316	346	374	400
1	100	141	173	200	224	245	265	283
	1	2	3	4	5	6	7	8
	Rate of Labor Input (*L*)							

between these extremes, such as b or c.[1] The concept of substitution is important because it means that managers can change the mix of capital and labor in response to changes in the relative prices of these inputs.

Second, in Table 6.1, if input rates are doubled, the output rate also doubles. For example, maximum production with one unit of capital and four units of labor is 200. Doubling the input rates to $K = 2$, $L = 8$ results in the rate of output doubling to $Q = 400$. The relationship between output change and proportionate changes in both inputs is referred to as *returns to scale.* In Table 6.1, production is characterized by constant returns to scale. This means that if both input rates increase by the same factor (e.g., both input rates double), the rate of output also will double. In other production functions, output may increase more or less than in proportion to changes in inputs. Returns to scale have implications for the size of individual firms and the number of firms in an industry. For example, in producing a product, if output increases more than in proportion to increases in inputs, that industry is likely to have only a few large firms. The U.S. automobile industry is an example. A more detailed discussion of the concept of returns to scale is included later in this chapter.

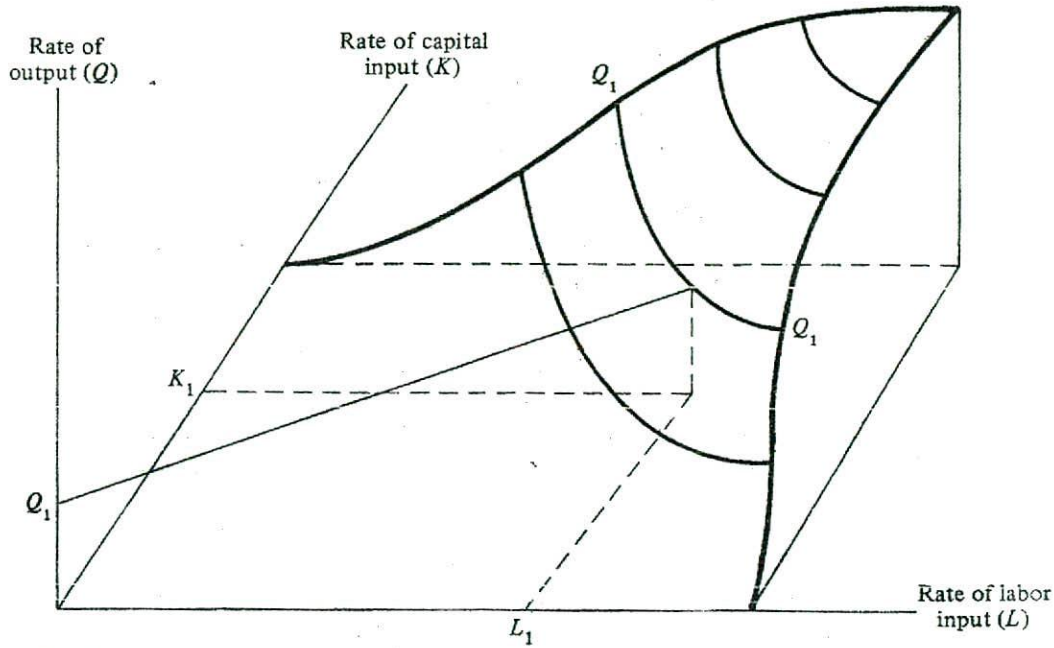
In contrast to the concept of returns to scale, when output changes because one input changes while the other remains constant, the changes in the output rates are referred to as *returns to a factor.* Note in the table that if the rate of one input is held constant while the other is increased, output increases but the successive increments become smaller. For example, from Table 6.1 it is seen that if the rate of capital input is held constant at 2 and labor is increased from $L = 1$ to $L = 6$, the successive increases in output are 59, 45, 38, 33, and 30. As discussed following, this relationship holds for virtually all production processes and is the basis for an important economic principle known as the law of diminishing marginal returns.

In Table 6.1, changes in output are shown only for discrete changes in the inputs. That is, only integer values of capital and labor are used. If the input rates of capital and labor can be varied continuously (i.e., any value of K and L such as $K = 6.24$ or $L = 3.15$ is possible), a production function traces a smooth, continuous surface, such as that shown in Figure 6.1. In this three-dimensional diagram, capital and labor are shown on the K and L axes and output is measured on the vertical axis, which is perpendicular to the K, L plane. The rate of output generated by a given capital–labor combination is found by identifying the point representing an input combination (such as K_1 and L_1 in Figure 6.1) and drawing a perpendicular line up to the production surface. The height of this perpendicular line defines the rate of output Q_1 corresponding to the input combination K_1 and L_1. In general, any point on the production surface defines the maximum output possible from the input combination associated with that point.

Although the production table provides considerable information on production possibilities, it does not allow for the determination of the profit-maximizing rate of output or even the best way to produce some specified rate of output. For example, the production data in Table 6.1 show four different combinations of capital and labor that will generate 245 units of output. Which is the best combination? Similarly, of the infinite number of possible output levels, which one will result in maximum profit for the

---

[1]The term *capital intensive* refers to a production system where the ratio of capital to labor is relatively high. In the *labor-intensive* case, the capital-to-labor ratio would be relatively low.



firm? The production function alone cannot answer these questions. As indicated previously, the production function, while a fundamental part of the decision-making process, is an engineering relationship and must be combined with data on the price of capital, labor, and output to determine the optimal allocation of resources in the production process.

Key Concepts

- The production function is an engineering concept that defines the maximum rate of output forthcoming from specified input rates of capital and labor.
- The Cobb-Douglas or multiplicative form of the production function, $Q = AK^{\alpha}L^{\beta}$, is widely used in economics because it accurately characterizes many production processes.
- Generally, a specified rate of output can be produced using different combinations of capital and labor. That is, there is substitutability between the factors of production.
- The concept of returns to scale refers to changes in production when all inputs are varied proportionately. Returns to a factor refers to changes in production associated with change in only one input.

PRODUCTION WITH ONE VARIABLE INPUT

The problem of optimal production will be approached in two ways. In this section it is assumed that the period of production is of such length that the rate of input of one factor of production is fixed. That is, the period is not long enough to change the input rate of that factor. The problem, then, is to determine the optimal rate of the variable input given the price of output, the price of the variable input, and the production technology as described by the production function. In the next section, both inputs will be allowed to vary and the optimal rates of both variable inputs, capital, and labor will be determined.

The period of time during which one of the inputs is fixed in amount is defined as the short run. In contrast, all inputs are variable in the long run. The period of time for the short run will vary among firms. For some firms, the short run may be a matter of days. For others, such as an electric utility company, the short run may be a number of years—the period of time necessary to plan and build a new generation unit.

Generally, at any point in time, the firm is operating in the short run. That is, the input rates of one or more factors are fixed. But most firms are continuously planning or considering changes in the entire scale of operation that would involve changes in all input rates. Thus, it is said that the firm plans in the long run but operates in the short run. For example, an automobile manufacturer may have six plants with maximum production capacity of 1.5 million vehicles per year. To build a new plant may take several years. At any particular time, the firm operates the existing plants—a short-run decision, but based on current and projected demand conditions, the firm will plan to either augment or reduce plant capacity in the future—a long-run decision.

The Product Functions

For a two-input production process, the total product of labor (TP_L) is defined as the maximum rate of output forthcoming from combining varying rates of labor input with a fixed capital input. Denoting the fixed capital input as $\bar{K}$, the total product of labor function

$$TP_L = f(\bar{K}, L) \tag{6-4}$$

Similarly, the total product of capital function is written as

$$TP_K = f(K, \bar{L}) \tag{6-5}$$

Two other product relations are relevant. First, marginal product (MP) is defined as the change in output per one-unit change in the variable input. Thus, the marginal product of labor is

$$MP_L = \frac{\Delta Q}{\Delta L}$$

and the marginal product of capital is

$$MP_K = \frac{\Delta Q}{\Delta K}$$

For infinitesimally small changes in the variable input, the marginal product function is the first derivative of the production function with respect to the variable input. For the general Cobb–Douglas production function,

$$Q = AK^{\alpha}L^{\beta}$$

the marginal products are

$$MP_K = \frac{dQ}{dK} = \alpha AK^{\alpha-1}L^{\beta}$$

and

$$MP_L = \frac{dQ}{dL} = \beta AK^{\alpha}L^{\beta-1}$$

Second, average product (AP) is total product per unit of the variable input and is found by dividing the rate of output by the rate of the variable input. The average product of labor function is

$$AP_L = \frac{TP_L}{L} \tag{6-6}$$

and the equation for the average product of capital is

$$AP_K = \frac{TP_K}{K} \tag{6-7}$$

Consider a hypothetical production function. If capital is fixed at two units, the rates of output generated by combining various levels of labor with two units of capital (i.e., the total product of labor) are as shown in Table 6.2. The average and marginal product of labor also are shown in the table.

The total product function can be thought of as a cross section or vertical slice of a three-dimension production surface, such as that shown in Figure 6.2a. Suppose the capital stock is fixed at K_1. The total product of labor function $f(K_1,L)$ is shown as the line starting at K_1 and extending through point a. Similarly, if the labor input is fixed at L_3, the total product of capital function is shown as the line beginning at L_3 and going through points a and b. Other total product functions are shown as the line beginning

Rate of Labor Input (L)	TP_L	AP_L	MP_L
0	0	—	—
1	20	20	20
2	50	25	30
3	90	30	40
4	120	30	30
5	140	28	20
6	150	25	10
7	155	22	5
8	150	19	−5

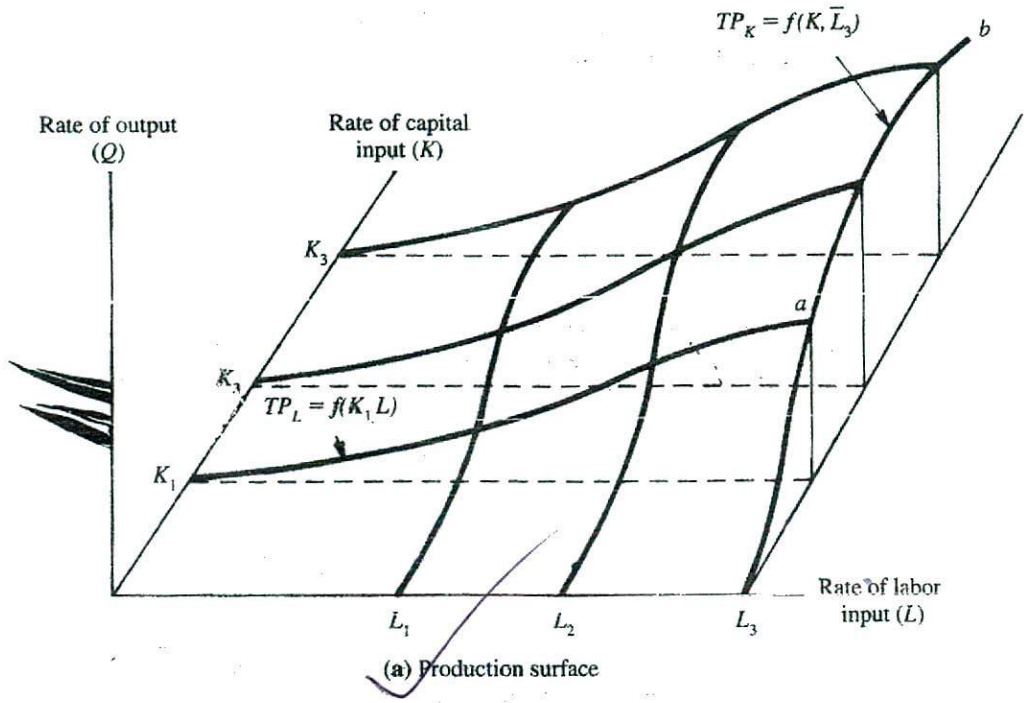


FIGURE 6.2 Production Surface and Total Product Function

at L_1, L_2, K_2, and K_3. If the cross section associated with $TP_K = f(K, L_3)$ was shown in a two-dimensional graph having output on the vertical axis and capital on the horizontal axis, it would appear as shown in Figure 6.2b.

Diminishing Marginal Returns

Consider a clothing manufacturer who has a 5,000-square-foot building housing 100 sewing machines. Obviously, having only one or two workers in such a plant would be inefficient. As more labor is added, production should increase rapidly as more machines are placed in operation and better coordination is achieved among workers and machines. However, as even more labor is added, the efficiency gains will slow and output will increase, but at a slower rate (i.e., marginal product will decline). Finally, a point may be reached where adding more labor actually will cause a reduction in total output, that is, where marginal product becomes negative. Conceivably, because only so many workers can be put in a finite space, enough labor could be added so that the production process would come to a standstill, reducing output to zero.

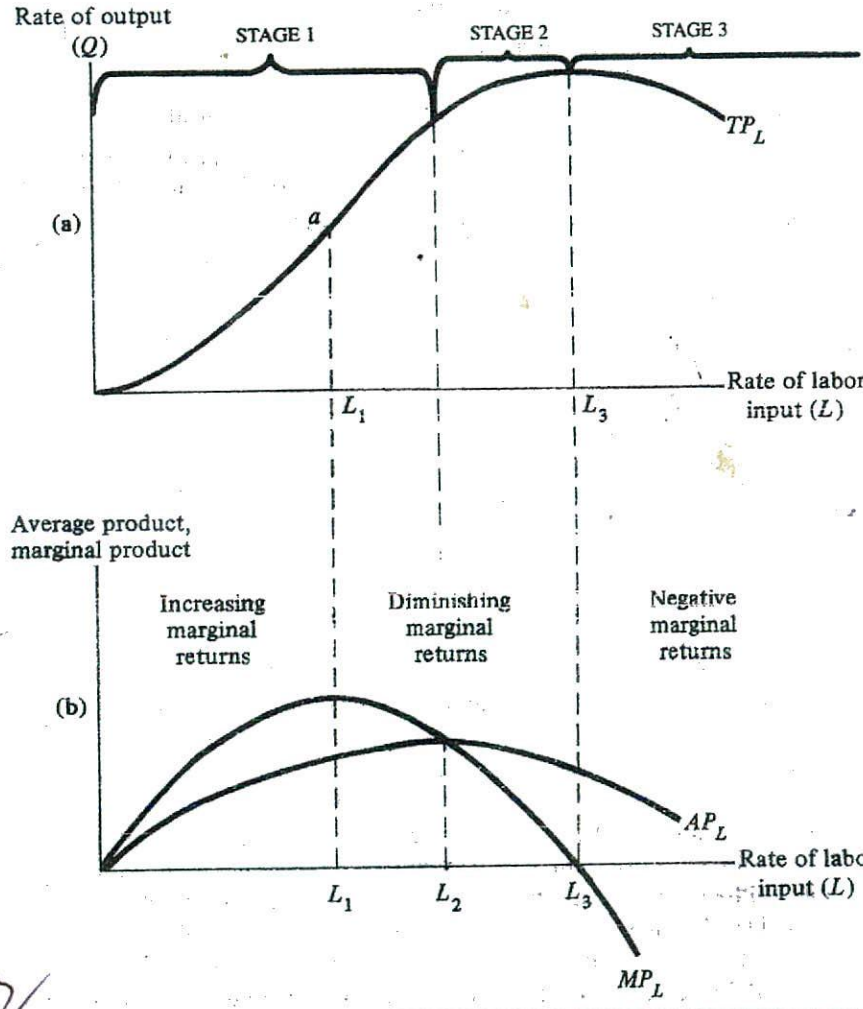
This example illustrates an important economic principle known as the *law of diminishing marginal returns*. This law states that when increasing amounts of the variable input are combined with a fixed level of another input, a point will be reached where the marginal product of the variable input will decline. This law does not result from a theoretical argument but is based on actual observation of many production processes. Virtually all studies of production systems have verified the existence of diminishing marginal returns.

Relationships among the Product Functions

A set of typical total, average, and marginal product functions for labor is shown in Figure 6.3. Total product begins at the origin, increases at an increasing rate over the range 0 to L_1, and then increases at a decreasing rate. Beyond L_3, total product actually declines. The explanation is as follows. Initially, the input proportions are inefficient—there is too much of the fixed factor, capital. As the labor input is increased from 0 to L_1, output rises more than in proportion to the increase in the labor input. That is, marginal product per unit of labor increases as a better balance of labor and capital inputs is achieved. As the labor input is increased beyond L_1, diminishing marginal returns set in and marginal product declines; the additional units of labor still result in an increase in output, but each increment to output is smaller. When the labor input has increased to L_3, total product reaches a maximum, and then, beyond L_3, the amount of labor has become excessive and slows the production process, with the result that total product actually declines.

Several relationships among the total, average, and marginal product functions are important:

1. Marginal product reaches a maximum at L_1, which corresponds to an inflection point (a) on the total product function. At the inflection point, the total product function changes from increasing at an increasing rate to increasing at a decreasing rate.
2. Marginal product intersects average product at the maximum point on the average product curve. This occurs at labor input rate L_2. Recall that whenever marginal product is above average product, the average is rising—it makes no difference whether marginal product is rising or falling. When marginal product is below average product, the average is falling. Therefore, the intersection must occur at the maximum point of average product.



3. Marginal product becomes negative at labor input rate L_3. This corresponds to the point where the total product curve reaches a maximum.

Production over the range 0–L_2 is defined as Stage 1. Here, the marginal product of labor is positive and increasing, but the marginal productivity of capital is actually negative. In this range, there is not enough labor to efficiently use the capital stock. Beyond L_3, the marginal product of labor is negative; here, there is too much labor combined with the capital stock. As it would make no sense to operate where the marginal product of either labor or capital is negative, the firm will not operate in either Stage 1 or 3. In Stage 2, the marginal product of both labor and capital is positive and declining. Production will occur only in this range.

Key Concepts

- The short run is that period of time for which the rate of input use of at least one factor of production is fixed. In the long run, the input rates of all factors are variable. The firm operates in the short run but plans in the long run.
- In the short run, total product is the set of output rates obtained by combining varying rates of one input with a fixed rate of the other input.
- Marginal product is the change in output associated with a one-unit change in the variable input (i.e., $MP_L = \Delta Q/\Delta L$) or the first derivative of the production function with respect to the variable input (i.e., $MP_L = dQ/dL$).
- Average product is the rate of output produced per unit of the variable input employed (i.e., $AP_L = Q/L$).
- The law of diminishing marginal returns states that when increasing rates of a variable input are combined with a fixed rate of another input, a point will be reached where marginal product will decline.

Optimal Employment of a Factor of Production

The General Motors Corporation has a worldwide physical capital stock valued at about $70 billion. Consider this to be the fixed input for the firm. About 760,000 workers are employed to use this capital stock. What principles guide the decisions about the level of employment? In general, to maximize profit, the firm should hire labor as long as the additional revenue associated with hiring of another unit of labor exceeds the cost of employing that unit. For example, suppose that the marginal product of an additional worker is two units of output (i.e., automobiles) and each unit of output is worth $20,000. Thus the additional revenue to the firm will be $40,000 if the worker is hired. If the additional cost of a worker (i.e., the wage rate) is $30,000, that worker will be hired because $10,000, the difference between additional revenue and additional cost, will be added to profit. However, if the wage rate is $45,000, the worker should not be hired because profit would be reduced by $5,000.

Formally stated, the basic principle is that additional units of the variable output should be hired until the marginal revenue product (MRP) of the last unit employed is equal to the cost of the input. The MRP is defined as marginal revenue times marginal product and represents the value of the extra unit of labor.[2] Thus labor is hired until MRP_L equals the wage rate (w):

$$MRP_L = w, \tag{6-8}$$

Similarly, if the labor input was fixed and the capital stock could be varied, capital would be employed until the marginal revenue product of capital equaled the price of capital (r), that is,

$$MRP_K = r \tag{6-9}$$

---

[2] In general, marginal revenue product is equal to $MR \cdot MP$. If price is constant $P = MR$ and marginal revenue product is $P \cdot MP$.

L	TP_L	MP_L	TR	MRP_L
0	0	—	$ 0	—
1	200	200	400	$400
2	283	83	566	166
3	346	63	692	126
4	400	54	800	108
5	447	47	894	94
6	490	43	980	86
7	529	39	1,058	78
8	565	36	1,130	72

Table 6.3 shows the total product, marginal product, total revenue, and marginal revenue product of labor for the production function $Q = 100\,K^{0.5}L^{0.5}$ and where the capital input rate has been fixed at four units and the price of output is $2. In the example, MRP_L can be determined either by multiplying each MP_L entry by $2 (the output price per unit) or by finding the change in total revenue for each one-unit increase in labor.

It is easily seen that the two methods are equivalent. Marginal revenue product is equal to marginal revenue multiplied by marginal product. That is,

$$MRP_L = MR \cdot MP_L$$

But

$$MR = \frac{\Delta TR}{\Delta Q}$$

and

$$MP_L = \frac{\Delta Q}{\Delta L}$$

Substituting, it is seen that

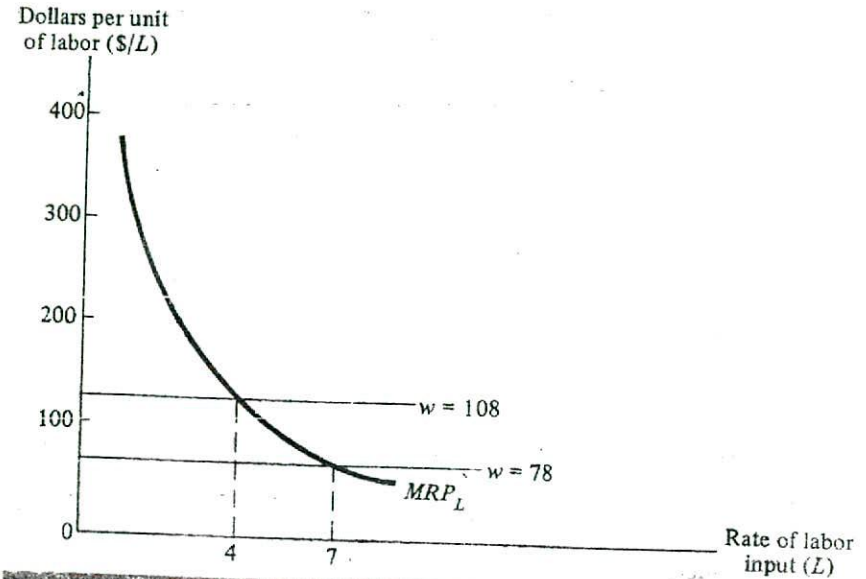
$$MRP_L = \frac{\Delta TR}{\Delta Q}\,\frac{\Delta Q}{\Delta L} = \frac{\Delta TR}{\Delta L}$$

Thus marginal revenue product can be written as the change in total revenue per one-unit change in the rate of labor input.

For infinitesimally small changes in the labor input, marginal revenue product is the first derivative of the total revenue function with respect to that input. That is,

$$MRP_L = \frac{d(TR)}{dL}$$

The optimal rate of labor to be hired depends on the wage rate. If a unit of labor costs $108, then four units of labor are hired because the firm will hire labor only as long



as MRP_L is greater than or equal to the wage rate. If the wage rate is lower, say \$78 per unit of labor, seven units will be hired. Clearly, if the wage rate is lower, more labor will be purchased.

The marginal revenue product is the labor demand function for the firm. That is, it indicates the amount of labor that will be hired at any wage rate. In graphing the labor demand curve, the vertical axis is measured in dollars, and the horizontal axis is measured as the rate of labor input. The labor demand curve is downward sloping because of the law of diminishing marginal returns.

The MRP_L curve corresponding to Table 6.3 is shown in Figure 6.4. The horizontal line at w = \$78 in the figure can be thought of as the supply function for labor facing the firm. This horizontal supply curve means that the firm can hire all the labor it wants at \$78 per unit. The optimum quantity of labor is determined by finding the intersection of the demand and supply functions, that is, the point where $MRP_L = w$. The figure shows that seven units of labor should be hired. If the wage rate increased to \$108 (shown by the w = 108 line in Figure 6.4), the quantity of labor demanded by the firm would fall to four units. If the wage rate is higher than \$108, less labor would be hired as the firm moved up the MRP_L curve.

Key Concepts

- Marginal revenue product (*MRP*) is found by multiplying the marginal product function by marginal revenue (i.e., $MRP = MR \cdot MP$).
- The marginal revenue product function for a productive factor is the demand curve for that factor.
- Additional units of a productive factor should be hired until the value of the marginal product of the input is equal to the price of that input.

Example **The Optimal Labor Input Rate**

The production function for Global Electronics is

$$Q = 2K^{0.5}L^{0.5}$$

with marginal product functions for labor and capital given by

$$MP_L = \frac{dQ}{dL} = 2\left(\frac{1}{2}\right)K^{0.5}L^{0.5-1} = \frac{K^{0.5}}{L^{0.5}} \text{ or } \frac{\sqrt{K}}{\sqrt{L}}$$

and

$$MP_K = \frac{dQ}{dK} = 2\left(\frac{1}{2}\right)K^{0.5-1}L^{0.5} = \frac{L^{0.5}}{K^{0.5}} \text{ or } \frac{\sqrt{L}}{\sqrt{K}}$$

respectively. Assume that the capital stock if fixed at nine units (i.e., $K = 9$). If the price of output (P) is \$6 per unit and the wage rate (w) is \$2 per unit, determine the optimal or profit-maximizing rate of labor to be hired. What labor rate is optimal if the wage rate increased to \$3 per unit?

Solution First, determine the MRP_L assuming that K is fixed at 9 (note that $P = MR$):

$$MRP_L = P \cdot MP_L = P\frac{\sqrt{K}}{\sqrt{L}} = 6\left(\frac{\sqrt{9}}{\sqrt{L}}\right) = \frac{18}{\sqrt{L}}$$

Now, equate the MRP_L function and the wage rate and solve for L. That is, set

$$MRP_L = w$$

and substitute, yielding

$$\frac{18}{\sqrt{L}} = 2 \text{ or } L = 81$$

Therefore, 81 units of labor should be employed.

If the wage rate increases to \$3 per unit of labor, the profit-maximizing condition $MRP_L = w$ would be

$$\frac{18}{\sqrt{L}} = 3$$
$$L = 36$$

This example shows that as the price of labor increases, the firm demands less labor. That is, the labor demand curve is downward sloping.

PRODUCTION WITH TWO VARIABLE INPUTS

If both capital and labor inputs are variable, a different set of analytical techniques must be applied to determine optimal input rates. There are three ways the firm may approach the problem of efficient resource allocation in production. They are (1) maximize production for a given dollar outlay on labor and capital, (2) minimize the dollar outlay on labor and capital inputs necessary to produce a specified rate of output, or (3) produce

the output rate that maximizes profit. For the profit maximization case, it will be shown that for each input, the marginal revenue product will equal the input price.

The first two problems are called *constrained optimization problems.* In problem (1), the constraint is a fixed-dollar outlay for capital and labor. In problem (2), the constraint is a specified rate of output that must be produced. However, in problem (3), the firm seeks that output level that will maximize profit; there is no constraint on either the budget available for production or the output level to be produced. The firm is only constrained by the limits set by the production function itself.

In this section, the approach to solving each of these problems is presented. A standard managerial economics technique using the concept of production isoquants and production isocosts is used to determine efficient input rate combinations for given production rates.

The Production Isoquant

In Figure 6.5, the three-dimensional production surface for the production function $Q = 100K^{0.5}L^{0.5}$ is shown. Think of an output rate, say $Q_1 = 490$, being specified and a "horizontal slice" cut through the production surface at that height. By cutting through the surface horizontally, the rate of output is held constant. This slice, denoted as Q_1Q_1, is a smooth curve that defines all combinations of capital and labor that yield a maximum production rate of 490. Two points on that isoquant, *a* and *b*, are used to show how the capital and labor input combinations are determined. Starting at point *a*, draw a perpendicular line from *a* to point *a'* on the capital–labor plane (i.e., the base of the diagram). Point *a'* denotes a capital input of 4 and labor input of 6. Repeating that process at *b* yields another capital–labor combination (3, 8) that also generates 490 units of output. If this process were repeated many times, it would trace out the smooth curve shown as the dashed curve $Q_1'Q_1'$ in the capital–labor plane. That curve is shown as the $Q = 490$ isoquant in the two-dimensional diagram in Figure 6.6.

Formally, an isoquant is the set of all combinations of capital and labor that yield a given output level. If fractional input units are allowed, there are an infinite number of points on any isoquant. Further, there is an isoquant through every point in the capital–labor space. Equivalently, there is an output rate corresponding to every combination of input rates. This implies that there are an infinite number of isoquants. For example, in addition to the isoquant for 490 units of output, isoquants for $Q = 200$ and $Q = 346$ are shown in Figure 6.6.

In general, isoquants are determined in the following way. First, a rate of output, say Q_0, is specified. Then, the production function is written as

$$Q_0 = f(K, L) \tag{6-10}$$

The combinations of K and L that satisfy this equation define the isoquant for output rate Q_0.

The slope of the isoquant shows the rate at which one input can be substituted for the other such that the level of output remains constant. This slope is referred to as the *marginal rate of technical substitution* (*MRTS*). For example, consider points *c*, *d*, and *e* on the 200-unit isoquant in Figure 6.6. Moving from point *c* to *d* involves substituting one additional unit of labor for two units of capital. That is, the marginal rate of substitution of labor for capital averages 1:2 over the range *c* to *d*. From point *d* to point *e*, it takes two units

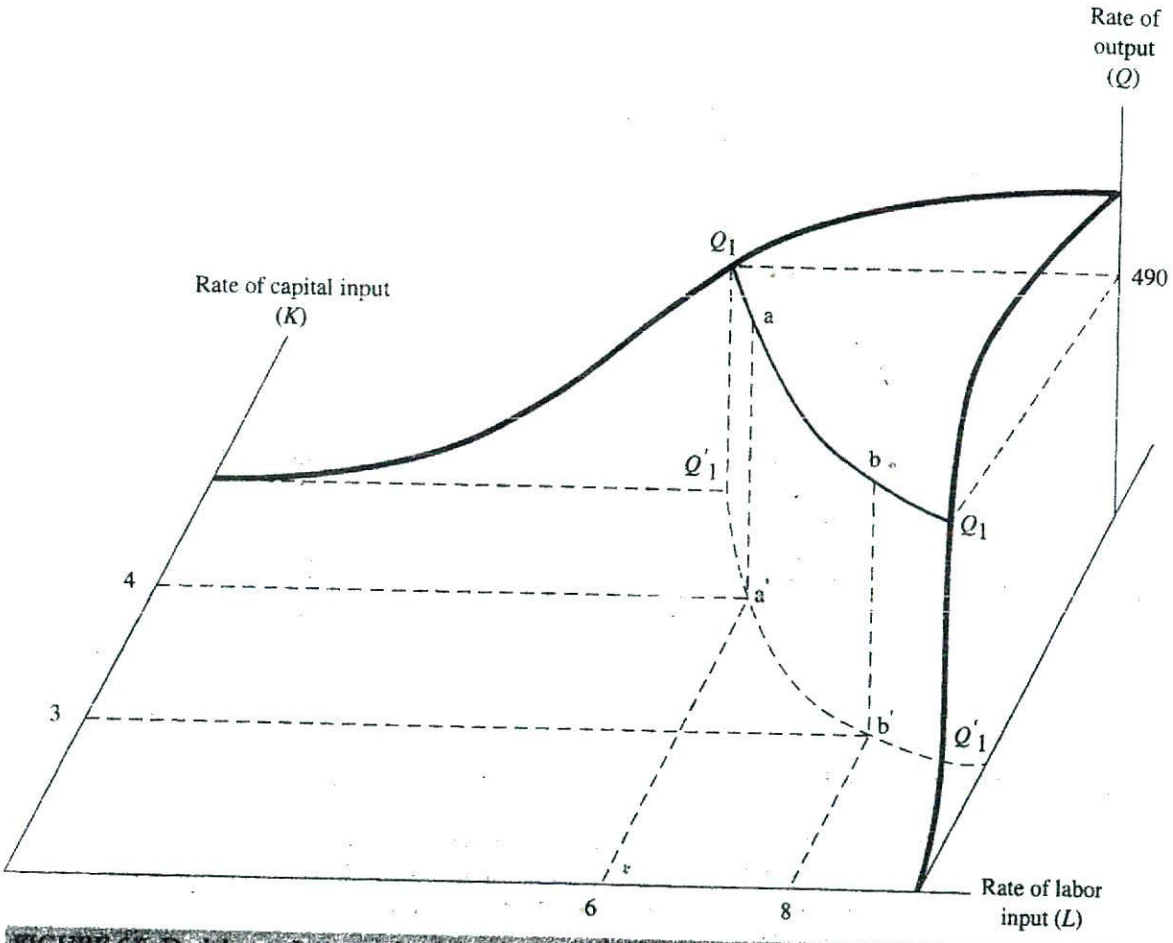
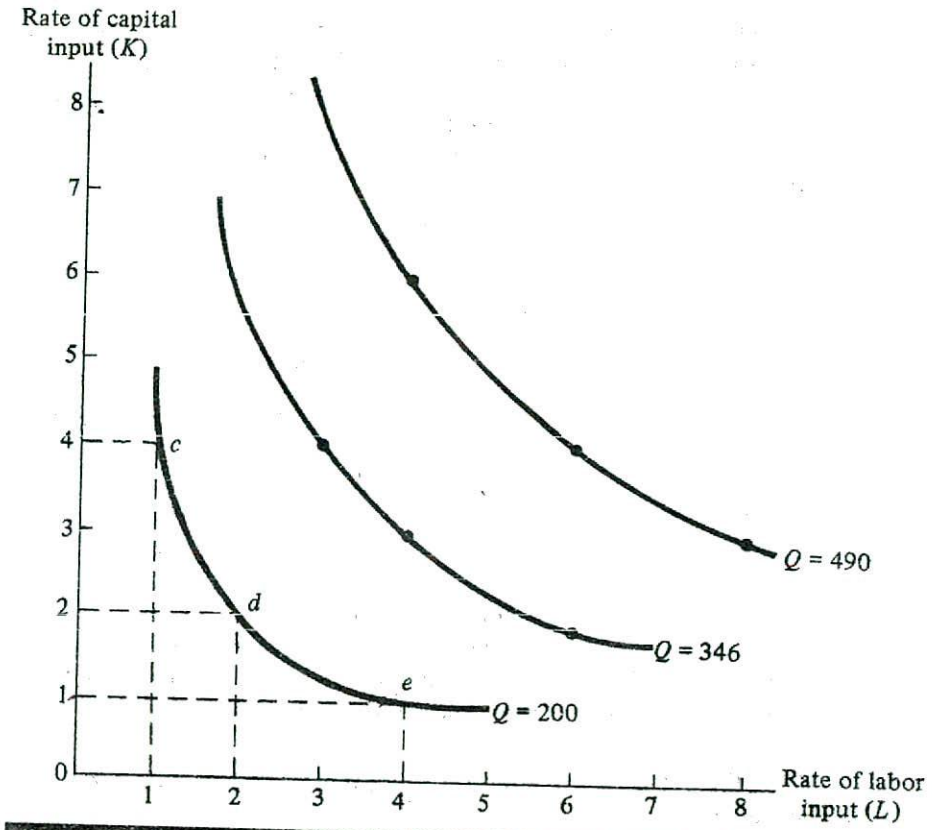


FIGURE 6.6 Deriving an Isoquant from the Production Surface

of labor to replace one unit of capital in order to maintain output at the 200 level. Thus, the marginal rate of substitution of labor for capital is 2:1 over the range *d* and *e*.

For most production functions, the isoquant is a smooth curve that is convex to the origin, as shown in Figure 6.6. The shape of this isoquant implies that inputs are imperfectly substitutable, and the rate of substitution declines as one input is substituted for another. For example, consider a factory having many machines but few workers. Adding an additional worker and reducing the number of machines may result in a much more efficient production system—that is, a relatively large increase in output would be obtained by adding that worker. But as more labor is added and machines removed, those efficiency gains fall. Thus it takes more and more workers to replace each machine.

It can be shown that the slope of an isoquant (i.e., the *MRTS*) is equal to the negative of the ratio of the marginal products, that is,



$$MRTS = -\frac{MP_L}{MP_K} \tag{6-11}$$

This relationship will be useful later on in determining the optimal rates of capital and labor to be hired when both inputs are variable.

Key Concepts

- An isoquant shows all combinations of capital and labor inputs that will produce a specified rate of output.
- The slope of an isoquant is the marginal rate of technical substitution or the rate that one input can be substituted for another so that a given rate of output is maintained.
- In general, the marginal rate of substitution diminishes as increased use of one input and less of another are combined.
- The marginal rate of technical substitution is equal to the negative of the ratio of the marginal products of the two inputs. That is, $MRTS = -(MP_L/MP_K)$.

The Production Isocost

The isoquant is a physical relationship that denotes different ways to produce a given rate of output. The next step toward determining the optimal combination of capital and labor is to add information on the cost of those inputs. This cost information is introduced by a function called a production isocost.

Given the per-unit prices of capital (r) and labor (w), the total expenditure (C) on capital and labor input is

$$C = rK + wL \tag{6-12}$$

For example, if $r = 3$ and $w = 2$, the combination of 10 units of capital and five units of labor will cost \$40. That is,

$$40 = 3(10) + 2(5)$$

For any given cost, C_0, the isocost line defines all combinations of capital and labor inputs that can be purchased for C_0.

Rewrite equation (6-12) by solving for K as a function of L,

$$K = \frac{C_0}{r} - \frac{w}{r}L \tag{6-13}$$

Equation (6-13) is an equation for a straight line where C_0/r is the vertical intercept and $-w/r$ is the slope. The ratio $-r/w$ is the rate that labor can be exchanged for capital in the market. For example, if $w = 2$ and $r = 3$, one unit of capital can be traded for 1.5 units of labor or one unit of labor can be traded for $\frac{2}{3}$ unit of capital.

Using the data from the preceding example ($w = 2, r = 3$, and $C = 40$), the isocost line becomes

$$40 = 3K + 2L$$

Solving for K yields

$$K = \frac{40}{3} - \frac{2}{3}L$$

or

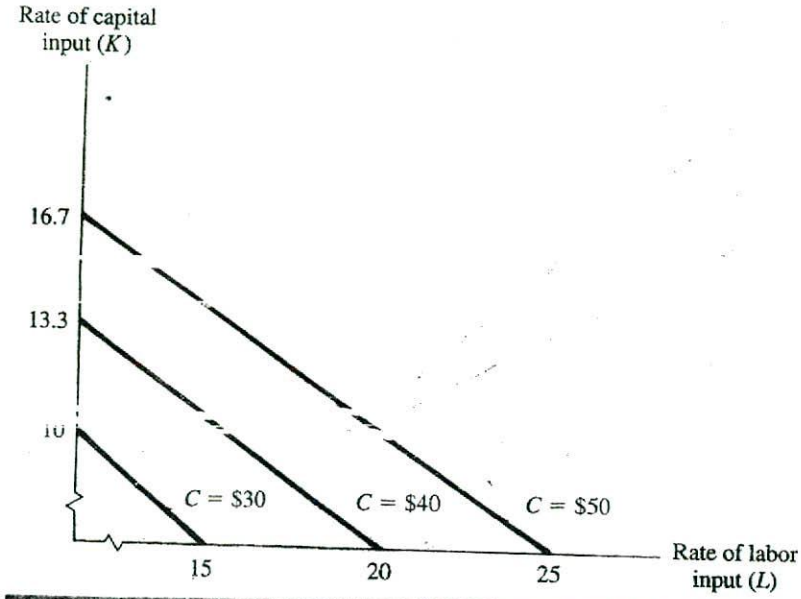
$$K = 13.33 - \frac{2}{3}L$$

This information is shown as the \$40 isocost line in Figure 6.7. Note the intercept points on the capital and labor axes. If all of the \$40 budget is spent on capital, 13.3 units can be purchased. Conversely, if the budget is spent entirely on labor, 20 units can be obtained.

If the budget constraint is increased to, say, \$50, the equation for the isocost equation becomes

$$50 = 3K + 2L$$

or



$$K = \frac{50}{3} - \frac{2}{3}L$$

Note that the intercept term on the capital axis has increased but the slope $(-\frac{2}{3})$ remains the same because the input prices are unchanged. That is, the new isocost has shifted outward but remains parallel to the \$40 isocost. The capital and labor intercepts (K, L) are now $(16.7, 0)$ and $(0, 25)$. Similarly, a decrease in the budget from \$40 to, say, \$30 causes a parallel shift in the function toward the origin. These three isocost lines (\$30, \$40, and \$50) are shown in Figure 6.7.

Now, consider how the isocost function shifts if an input price changes instead of the budget amount. Initially, assume that $C_0 = 40$, $w = 5$, and that r is variable. Thus the equation for the isocost is

$$K = \frac{40}{r} - \frac{5}{r}L$$

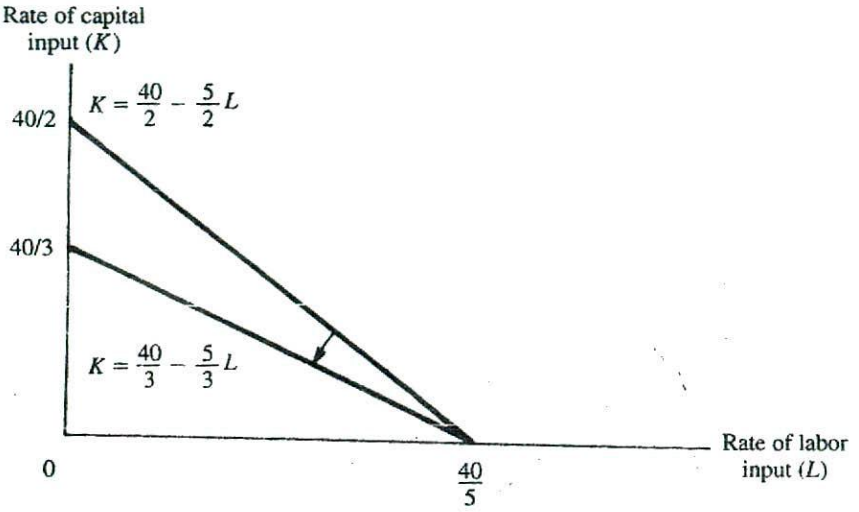
If r increases from, say, 2 to 3, the equation for the isocost changes from

$$K = \frac{40}{2} - \frac{5}{2}L$$

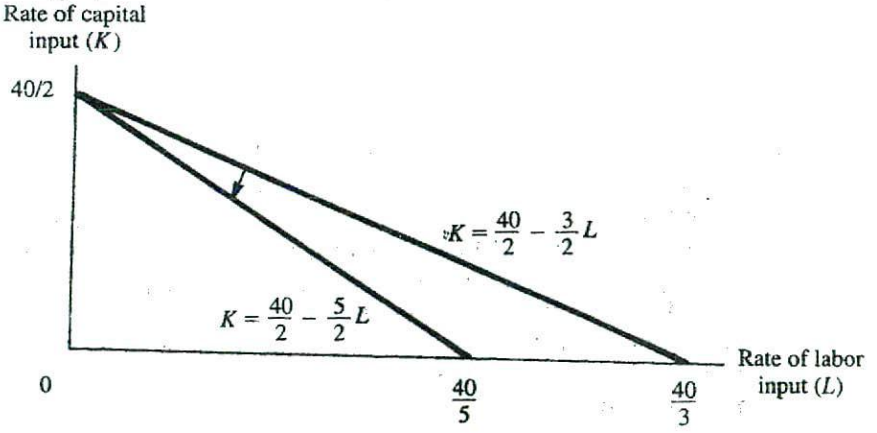
to

$$K = \frac{40}{3} - \frac{5}{3}L$$

These isocosts are shown in Figure 6.8a. Note that both the vertical (capital) intercept and the slope of the isocost function have changed, but the horizontal or labor intercept



(a) Production isocosts for different prices of capital



(b) Production isocosts for different prices of labor

FIGURE 6.8 Production Isocosts for Different Prices of Capital and Labor

is unchanged. That intercept is determined by the ratio C_0/w, which is not influenced by a change in the price of capital. Thus, as r changes, the isocost pivots about the point (40/5, 0), which is the horizontal intercept.

Conversely, if the price of capital is held constant at 2 and the price of labor changes from, say, 3 to 5, the \$40 isocost function will pivot about the point (40/2) on the vertical axis, as shown in Figure 6.8b.

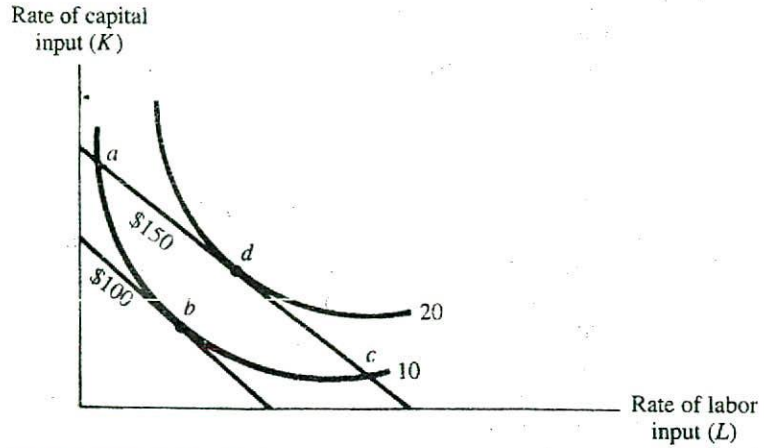


FIGURE 6.9 Using Isocost and Isoquant Functions to Solve the Production Problem

Key Concepts

- The isocost function is the set of all combinations of capital and labor that can be purchased for a specified total cost.
- Changes in the budget amount, C_0, cause the isocost line to shift in a parallel manner. Changes in either the price of labor or capital cause both the slope and one intercept of the isocost function to change.

Optimal Employment of Two Inputs

When both capital and labor are variable, determining the optimal input rates of capital and labor requires that the technical information from the production function (i.e., the isoquants) be combined with the market data on input prices (i.e., the isocost functions).

Consider the problem of minimizing the cost of a given rate of output. Specifically, suppose that the firm's objective is to produce 10 units of output at minimum cost. To help analyze this problem, two production isoquants are shown in Figure 6.9. The infinite number of capital–labor combinations that yield an output of 10 are indicated by the 10-unit isoquant. Three of these combinations are indicated by points *a*, *b*, and *c*. Points *a* and *c* are on the \$150 isocost and *b* is on the \$100 isocost. Of these, clearly *b* is the best of the three in the sense of being the lowest cost. In fact, *b* is the absolute minimum cost combination of capital and labor. At point *b*, the 10-unit isoquant is tangent to the \$100 isocost line. Note that all other input combinations shown on the 10-unit isoquant would correspond to higher isocost curves, thus costing more than \$100. That is, any other capital–labor combination on the 10-unit isoquant will be on a higher isocost line. Furthermore, there is no input combination that costs less than \$100 that will produce 10 units of output.

At the tangency of the 10-unit isoquant and the $100 isocost, the slopes of the two functions are equal. Thus, the marginal rate of technical substitution (i.e., the slope of the isoquant) equals the price of labor divided by the price of capital. That is,

$$MRTS = \frac{w}{r} \tag{6-14}$$

Equation (6-14) is a necessary condition for efficient production. If this equality does not hold (such as at points *a* and *c*), there is some other combination of capital and labor inputs that will reduce the cost of producing 10 units of output. Equivalently, there is a way to move along the 10-unit isoquant to a lower isocost.

Consider a different production problem. Suppose that the objective is to maximize output given a budget constraint of $150. Now the choice of input combinations is limited to points on the $150 isocost function. In Figure 6.9 three points are shown on that isocost line, *a*, *d*, and *c*. These all satisfy the budget constraint, but they are on different isoquants. For example, output at points *a* and *c* is 10 units, but at point *d* output is 20 units. Clearly, *d* is the preferred point among these three because a higher rate of output is produced. In fact, *d* is the best of all points on the $150 isocost. No other combination of capital and labor that costs $150 will yield as much output. Again, at that optimal point, the isoquant is tangent to the isocost. Hence the same efficiency condition ($MRTS = w/r$) applies. That is, the marginal rate of technical substitution in production must equal the rate of exchange of labor for capital in the market.

Regardless of the production objective, efficient production requires that the isoquant be tangent to the isocost function. If the problem is to maximize output subject to a given cost, the solution is found by moving along the specified isocost until the tangency is found. If the problem is to minimize cost subject to an output constraint, the solution is found by moving along the specified isoquant until the tangency is found. The same efficiency rule holds in both problems. For example, consider point *d* in Figure 6.9. That point can be thought of either as the minimum-cost capital–labor combination for producing 20 units of output or as the point of maximum output obtainable on a $150 budget.

These principles can be used to test for efficient resource allocation in production. It has been shown that the slope of the isocost is the negative of the ratio of the wage rate and price of capital (i.e., $-w/r$) and that the slope of the isoquant is the negative of the ratio of the marginal product of labor to that of capital (i.e., $-MP_L/MP_K$). Further, it has been shown that at a point of tangency, the slopes of both the isocost and isoquant are equal. Thus

$$-\frac{MP_L}{MP_K} = -\frac{w}{r}$$

or

$$\frac{MP_L}{MP_K} = \frac{w}{r}$$

This condition must be met for efficient production.

Rewriting this efficiency condition as

$$\frac{MP_L}{w} = \frac{MP_K}{r} \tag{6-15}$$

suggests an important principle. For an input combination to be efficient, the marginal product per dollar of input cost must be the same for both inputs.[3] For example, consider the following inefficient situation. Assume that both w and r are equal to 2 but that the marginal product of labor is 10, compared to 8 for capital. Thus we have

$$\frac{10}{2} > \frac{8}{2}$$

This cannot be an efficient input combination. The firm is getting more output per dollar for funds spent on labor than on capital. But because the input prices are the same, labor can be traded for capital on a one-to-one basis. If one unit of capital is sold to obtain one unit of labor, the reduction of capital by one unit causes output to fall by 8, but increasing the labor input by one unit will increase output by 10. Thus the substitution of labor for capital would result in a net increase of two units of output at no additional cost.

The inefficient combination corresponds to a point such as a in Figure 6.9. At that point the ratio of capital to labor is too high. The profit-maximizing firm will substitute labor for capital by moving down the isocost line. Conversely, at a point such as c in Figure 6.9, the reverse is true—there is too much labor, and the inequality

$$\frac{MP_L}{r} < \frac{MP_K}{r}$$

will hold. That is, the firm generates more output per dollar spent on capital than from dollars spent on labor. Thus the firm should substitute capital for labor.

Suppose that the firm is producing at the inefficient point c in Figure 6.9. If the problem is to minimize the cost of producing a given rate of output, the firm would move from point c along the 10-unit isoquant to b, thereby reducing cost by \$50 while maintaining the rate of production at 10 units. Alternatively, if the firm is maximizing output subject to a \$150 cost constraint, it would move from c along the \$150 isocost to point d, where that isocost is tangent to the 20-unit isoquant. Note that in the latter case, output would increase from 10 to 20 at no additional cost.

Profit Maximization

The efficiency condition just discussed is necessary but not sufficient for profit maximization. There are many points of tangency between isocosts and isoquants. For example, both points b and d in Figure 6.9 are efficient resource combinations. However, profits will be different at each point. Thus, the problem is to determine that one point among the many efficient points that results in the largest profit. That the efficiency condition is necessary for maximum profit is obvious. If the firm is not operating at an efficient point, there will be some way to reduce the cost of that level of output and thereby increase profit. Thus, only the efficient points defined by the tangencies of isocost and isoquant functions need to be considered.

To maximize profit, it is necessary to fall back on the rule for an efficient input rate when only one input is variable. That is, both inputs must be hired until the marginal revenue product equals the price of the input for both capital and labor. That is, the conditions for profit maximization are that

[3] This principle extends directly to production functions with more than two inputs. In general, efficient production requires that the ratio of marginal product to input price be equal for all inputs.

$$MRP_K = r \tag{6-16}$$

and

$$MRP_L = w \tag{6-17}$$

Marginal revenue product measures the additional dollars of revenue added by using one more unit of input. As long as *MRP* is greater than the cost of the input, profit can be increased by adding more of that input. That is, if an additional unit of an input adds more to revenue than that input cost, profits will increase, and those profits will increase until marginal revenue product equals the input price.

These conditions for a profit maximization imply that the condition for efficient production (i.e., $MP_L/P_L = MP_K/P_K$) will be met. This is shown by rewriting equations (6-16) and (6-17), that is,

$$MRP_K = r$$

and

$$MRP_L = w$$

as

$$P \cdot MP_L = w \tag{6-18}$$

and

$$P \cdot MP_K = r \tag{6-19}$$

Dividing equation (6-18) by (6-19) yields

$$\frac{MP_L}{MP_K} = \frac{w}{r}$$

and rewriting results in the efficiency condition

$$\frac{MP_L}{w} = \frac{MP_K}{r}$$

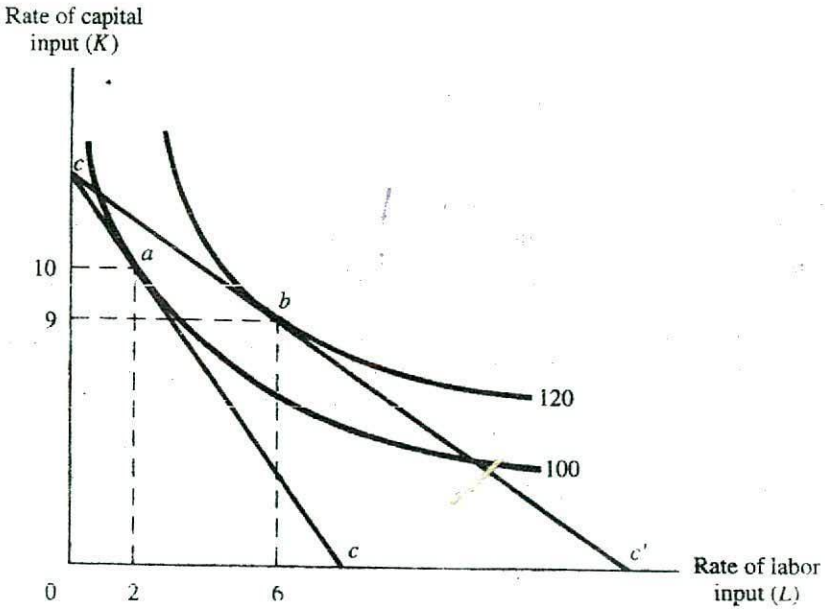
Therefore, if a firm is maximizing profit, it follows that it must be operating efficiently.

Key Concepts

- Efficient production requires that the isoquant function be tangent to the isocost function. At those points, the marginal product per dollar of input cost is equal for both inputs. That is,

$$\frac{MP_L}{w} = \frac{MP_K}{r}$$

- If the condition for efficient production is not met, there is some way to substitute one input for the other that will result in an increase in production at no change in total cost.
- Profit maximization requires that inputs be hired until $MRP_K = r$ and $MRL_L = w$. The conditions also imply that $MP_L/w = MP_K/r$.



Changes in Input Prices

If the price of one input, say labor, increases, the firm will adjust the input mix by substituting capital for labor. If the price of labor declines, thus making labor relatively less expensive, labor will be substituted for capital. In general, if the relative prices of inputs change, managers will respond by substituting the input that has become relatively less expensive for the input that has become relatively more expensive.

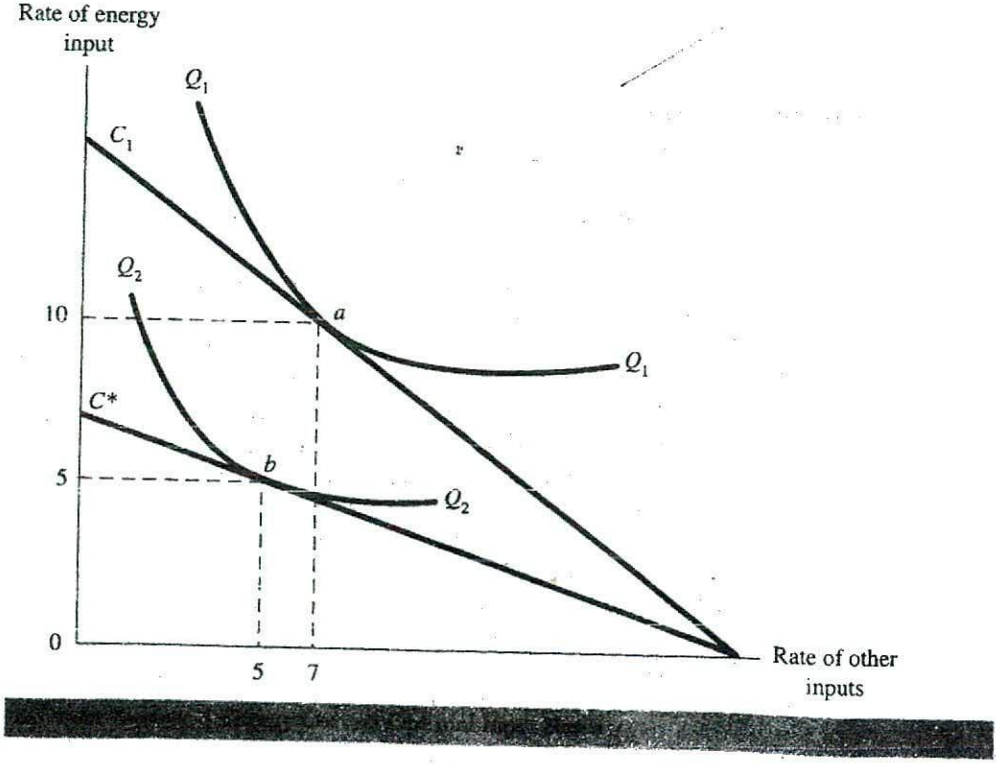
The isoquant–isocost framework can be used to demonstrate this principle. Consider Figure 6.10 on page 207. Suppose the firm currently is operating at point *a* where 100 units of output are produced using the resource combination ($K = 10, L = 2$). This is an efficient resource mix because the 100-unit isoquant is tangent to the isocost line *cc* at point *a*. Assume that the firm's goal is to maximize production subject to a cost constraint (i.e., the firm is limited to resource combinations on a given isocost function).

Now, assume that the price of labor falls while the price of capital remains unchanged (i.e., labor has become relatively less expensive). The isocost pivots to the right from *cc* to the isocost *cc'*. The reduction in the price of labor means that the firm is able to increase the rate of production. Hence the firm moves from point *a* to point *b*, which is a new efficient resource combination. That is, the new isocost is tangent to the 120-unit isoquant at point *b*. Now nine units of capital and six units of labor are employed. Note that at point *a*, the efficient ratio of capital to labor was 5:1. Now the efficient ratio of the two inputs is 3:2. The reduction in the price of labor has caused the firm to substitute that relatively less expensive input for capital.

Case Study

Input Substitution in Response to Higher Energy Prices

The decade of the 1970s was one of rapidly rising prices for virtually all energy products. The price of gasoline, fuel oil, and natural gas increased much more rapidly than the prices of most other products and services. For example, during the period 1971–1980, the real price (i.e., the price adjusted for overall inflation) of crude oil, natural gas, and coal increased 240 percent, 347 percent, and 113 percent, respectively. A major reason for these price increases was the reduction in output orchestrated by the Organization of Petroleum Exporting Countries (OPEC), consisting of Venezuela, Nigeria, and all of the oil-producing countries in the Middle East. Collectively, this group accounted for more than one-half of world oil production, and, by restricting supply, it was able to cause significant price increases.



Energy Consumption per Dollar of Value Added in Selected Industries

	Sector					
Year	*All Manufacturing*	*Paper*	*Organic Chemicals*	*Petroleum Refining*	*Steel*	*Aluminum*
1971	52.5	316.2	277.9	631.4	314.7	418.5
1977	42.3	308.7	193.9	573.4	282.7	379.9
Percent change	−19.4	−2.4	−30.2	−9.2	−10.2	−9.2

Source: U.S. Department of Commerce, Bureau of the Census, *Statistical Abstract of the United States: 1981* (Washington, D.C.: U.S. Government Printing Office, 1981).

Because energy is an important input in many production systems, principles of managerial economics predict that firms will substitute other inputs for the relatively more expensive energy products. In the figure, the input of energy is measured on the vertical axis, and a composite measure of other inputs is measured on the horizontal axis. Assume that before the increase in energy prices, a hypothetical firm is producing at point *a*, where the Q_1 isoquant is tangent to the C_1 isocost. The optimal input ratio is 10:7 or 10 units of energy for every 7 units of the "other input."

If the price of energy increased, the isocost would pivot downward from C_1 to C^*. The firm now will operate at point *b*, where the ratio of energy to other inputs is 5:5, or one unit of energy for every unit of the "other" input. Thus, the result of the higher price of energy is that the firm has substituted other inputs for energy.

As shown in the preceding table, producers in the United States did reduce their dependence on energy by substituting other inputs for energy. As measured by energy consumption (thousand Btu) per dollar of value added, dependence on this input was reduced significantly. Even an energy-producing sector, petroleum refining, conserved on its use of energy by using relatively more of other inputs. Overall, U.S. energy consumption in BTUs per dollar of gross domestic product fell almost 30 percent in the period 1970–1985. Since then, real energy prices have actually fallen, and there has been relatively little change in that measure. ■

The Expansion Path

Consider the system of isoquants and isocosts shown in Figure 6.11. Suppose that a firm is producing 1,000 units of output using 10 units of capital and 10 units of labor (i.e., point *a*) and the input prices are $w = 2$ and $r = 2$. Thus the cost of this input combination is $40. At point *a*, the 1,000-unit isoquant is tangent to the $40 isocost line. If output is to be increased, how much capital and labor will be hired? That is, how will the firm expand production? Clearly it will move to point *b* if 1,500 units are to be produced and then to point *c* if 1,750 units of output are to be produced. In general, the firm expands by moving from one tangency or efficient production point to another. These efficient points represent the expansion path. An expansion path is formally de-

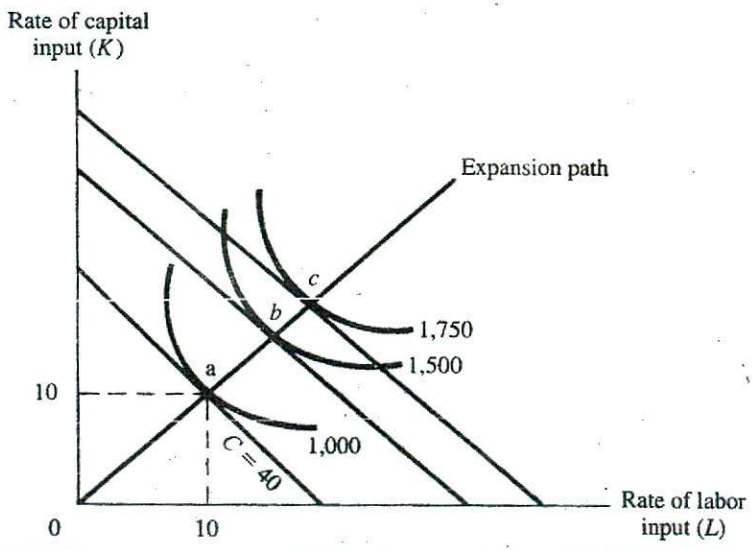


FIGURE 6.11 The Expansion Path for the Firm

fined as the set of combinations of capital and labor that meet the efficiency condition $MP_L/w = MP_K/r$.

An equation for the expansion path can be determined by first substituting the marginal product equations and input prices into the efficiency condition, and then by solving for capital as a function of labor. For example, suppose that the production function is of the form $Q = 100K^{0.5}L^{0.5}$. The corresponding marginal product functions are

$$MP_L = \frac{dQ}{dL} = 50\frac{K^{0.5}}{L^{0.5}}$$

and

$$MP_K = \frac{dQ}{dK} = 50\frac{L^{0.5}}{K^{0.5}}$$

Substituting the marginal product equations in the efficiency condition ($MP_L/MP_K = w/r$) yields

$$\frac{50\dfrac{K^{0.5}}{L^{0.5}}}{50\dfrac{L^{0.5}}{K^{0.5}}} = \frac{w}{r}$$

Solving for K gives:

$$K = \frac{w}{r}L \qquad (6\text{-}20)$$

This expression is the equation for the expansion path for the production function $Q = 100K^{0.5}L^{0.5}$. If w and r are known, equation (6-20) defines the efficient combination of capital and labor for producing any rate of output. That is, it is an equation for an expansion path such as that shown in Figure 6.11. For example, if $w = 1$ and $r = 1$, the expansion path defined by equation (6-20) would be

$$K = L$$

If $w = 2$ and $r = 1$, the equation for the expansion path would be

$$K = 2L$$

If the expansion path is known, then knowing the isoquant–isocost system is not necessary to determine efficient production points. The firm will only produce at those points on the expansion path.

The expansion path indicates optimal input combinations, but it does not indicate the specific rate of output associated with that rate of input use. The output rate is determined by substituting the equation for the expansion path into the original production function. In the example, substituting the equation for the expansion path, $K = (w/r)L$, into the production function, $Q = 100K^{0.5}L^{0.5}$ yields

$$Q = 100\left(\frac{w}{r}L\right)^{0.5} L^{0.5}$$

or

$$Q = 100L\left(\frac{w}{r}\right)^{0.5} \tag{6-21}$$

The two equations (6-20) and (6-21) have three unknowns: K, L, and Q. [Recall that the prices of labor (w) and capital (r) are assumed to be known.] If the value of K, L, or Q is given, the efficient rate of the other two variables can be calculated.

Consider the problem of determining the efficient input combination for producing 1,000 units of output if $w = 4$ and $r = 2$. The steps are as follows. First, substitute $Q = 1{,}000$, $w = \$4$, and $r = \$2$ into equation (6-21) and solve for L. Thus

$$1{,}000 = 100L\left(\frac{4}{2}\right)^{0.5}$$

$$L = 7.07$$

Then, substitute $L = 7.07$ into (6-20), the equation for the expansion path, to find K.

$$K = \left(\frac{4}{2}\right)7.07$$

$$K = 14.14$$

Thus the input combination ($K = 14.14$, $L = 7.07$) is the most efficient way to produce 1,000 units of output.

How would the input mix change if the price of capital increased to $r = 4$ and the firm still wanted to produce 1,000 units of output? Again, substitute $Q = 1{,}000$ into equation (6-21) and solve for L:

$$1{,}000 = 100L\sqrt{\frac{4}{4}}$$

$$L = 10$$

Now substitute $L = 10$ into equation (6-20) and solve for K:

$$K = \left(\frac{4}{4}\right)L$$

$$K = 10$$

The new efficient combination is ($K = 10$, $L = 10$). The firm responded to the higher price of capital by substituting labor for capital. That is, the capital input was reduced from 14.14 to 10, and the labor input was increased from 7.07 to 10.

Key Concepts

- If the price of one or both inputs changes, the firm responds by substituting the input that has become relatively less expensive for the other input.
- The firm expands production by moving from one efficient production point (where the isoquant and isocost functions are tangent) to another. These efficient points define the firm's expansion path.
- By using the production function and the production efficiency conditions together, the optimal level of capital and labor used to produce any rate of output can be determined. This expansion path determines the efficient input combinations for any output rate.

ECONOMIES OF SCALE AND SCOPE

In general, the cost of producing and marketing products depends both on the scale (i.e., the amount of labor and capital employed) and the scope (i.e., the array of different goods and services produced) of the firm's operations. The relationship of per-unit costs to changes in these two factors are referred to as economies of scale and economies of scope.

Economies of Scale

A given rate of input of capital and labor defines the scale of production. Proportionate changes in both inputs result in a change in that scale. The term *returns to scale* refers to the magnitude of the change in the rate of output relative to the change in scale. For example, given the production function

$$Q = 100K^{0.5}L^{0.5}$$

if both capital and labor are 10, output will be 1,000. A doubling of both input rates to 20 will result in a doubling of output to 2,000. This proportionate response of output to change in inputs is defined as *constant returns to scale.*

In general, there is no reason to expect that output will always change in proportion to the change in inputs. Output might increase more than in proportion (*increasing*

returns to scale) or less than in proportion (*decreasing returns to scale*). Returns to scale are formally classified as follows. Given the general production function

$$Q = f(K, L) \tag{6-22}$$

if both inputs are changed by some factor λ, output will change by a factor h. That is,

$$hQ = f(\lambda K, \lambda L) \tag{6-23}$$

If $h = \lambda$, the production function is said to be characterized by constant returns to scale because the change in output is proportional to the change in both inputs. If $h < \lambda$, there are decreasing returns to scale, and if $h > \lambda$, returns to scale are increasing.

There is a simple way to test for constant, decreasing, or increasing returns to scale. Solve the production function (i.e., determine the rate of output) for one set of input values, double both inputs, and again solve for output. If output doubled, the production function is characterized by constant returns to scale over that range of output. If output changed by less than twice the initial rate, decreasing returns to scale apply. Finally, if output more than doubled, the function exhibits increasing returns to scale.[4]

For production functions of the Cobb–Douglas type (i.e., $Q = AK^{\alpha}L^{\beta}$), the arithmetic sum of the exponents (i.e., $\alpha + \beta$) can be used to determine if returns to scale are decreasing, constant, or increasing. This is demonstrated by taking the basic Cobb–Douglas production function

$$Q = AK^{\alpha}L^{\beta}$$

and doubling both inputs, which will increase Q by a factor h, that is,

$$hQ = A(2K)^{\alpha}(2L)^{\beta}$$

Rewriting yields

$$hQ = 2^{\alpha}2^{\beta}(AK^{\alpha}L^{\beta})$$

or

$$hQ = 2^{\alpha+\beta}(AK^{\alpha}L^{\beta})$$

But $Q = AK^{\alpha}L^{\beta}$. Hence the factor $h = 2^{\alpha+\beta}$ and will be less than 2, equal to 2, or greater than 2, depending on whether $(\alpha + \beta)$ is less than 1, equal to 1, or greater than 1, respectively. Thus the three possibilities are

Sum of Exponents $(\alpha + \beta)$	*Returns to Scale*
Less than one	Decreasing
Equal to one	Constant
Greater than one	Increasing

For example, the production function $Q = 10K^{0.5}L^{0.6}$ is characterized by increasing returns to scale because $\alpha + \beta$ (i.e., 0.5 + 0.6) is greater than unity. In contrast, the function $Q = 20K^{0.4}L^{0.5}$ exhibits decreasing returns to scale because the sum of the exponents is 0.9.

---

[4]There are production functions that are characterized by increasing returns over part of the output range and decreasing returns over another part of the range. For these functions, this simple technique will test for the nature of returns to scale only for that part of the output range that is evaluated.

The principle applies to Cobb–Douglas production functions with any number of inputs. Given a production function of the form $Q = AK_1^{\alpha}K_2^{\beta}L_1^{\delta}L_2^{\nu}$, returns to scale would be decreasing, constant, or increasing if $\alpha + \beta + \delta + \nu$ is less than 1, equal to 1, or greater than 1, respectively.

Sources of Economies of Scale

There are several reasons why increasing returns to scale occur. First, technologies that are cost-effective at high levels of production generally have higher unit costs at lower levels of output. For example, the million-dollar machinery used for cutting and stamping auto bodies by General Motors would make little sense for use by a small custom car manufacturer. Geometric relations are another factor causing decreasing average costs. A gas company using 12-inch pipe has 3 cubic inches of pipe volume per square inch of pipe surface, while a firm with output sufficient to justify 24-inch pipe will have 6 cubic inches of volume per square inch of surface. The reduced cost per unit of pipe volume occurs because the materials requirement varies with the diameter of the pipe, while the volume varies with the square of the radius of the pipe. Thus the larger firm using bigger pipelines will have lower unit costs.

Two other causes of increasing returns are specialization of labor and inventory economies. As a firm becomes larger, the demand for employee expertise in specific areas grows. Instead of being generalists, workers can concentrate on learning all the aspects of particular segments of the production process. Usually, a worker who only has to perform one task can do it more rapidly and more accurately than one who must do many different jobs. Size also affects unit costs because larger firms may not need to increase inventories or replacement parts proportionately with size. For example, suppose that a small firm uses a machine that is critical to the firm's operations. Let the probability of a machine breakdown during a month be 0.10. If a replacement is not readily obtainable from a nearby supplier, the firm may be forced to keep a backup on the premises. Suppose that a larger firm uses five of the same machines, each with a 0.10 probability of malfunctioning. The probability that two of those machines will break down in a month is 0.1×0.1, or 0.01. The likelihood that the five machines will become inoperative is 0.1^5, which is essentially zero. Thus, whereas the small firm may be required to have one backup machine for its one operating machine, the larger firm may have a high degree of reliability with a much lower ratio of backup to operating machines and, hence, a lower per-unit cost for output.

Decreasing returns to scale may occur because the firm grows so large that management cannot effectively manage it. For example, the costs of gathering, organizing, and reviewing information on all aspects of a large firm may increase more rapidly than output. Further, managing large numbers of employees and coordinating the several divisions of a large firm may be difficult.

Transportation costs also may be a factor explaining decreasing returns to scale. If the firm consolidates two or more geographically dispersed plants, production costs may decline, but this decline could be offset by the higher average cost of shipping a unit of output to customers. The average distance shipped will be higher for one source than it will for two or more sources.

Finally, as plant size increases, the firm may employ such a large share of the local labor force that it will have to increase wage rates in order to attract more labor. These higher labor costs may offset other sources of cost reduction associated with the larger plant size.

Economies of Scope

Firms often find that per-unit costs are lower when two or more products are produced. Sometimes the firm will have excess capacity that can be used to produce other products with little or no increase in its capital costs. One example is the firm that reconfigured its passenger planes each night by removing seats in order to haul packages and freight. Ski resorts have developed various warm-weather activities (e.g., Alpine slides, mountain bike paths, etc.) to allow them to use ski lifts on a year-round basis.

Other firms have taken advantage of their unique skills or comparative advantage in marketing to develop products that are complementary with the firm's existing products or that would simply be logical items for the firm's sales staff to sell on their regular calls on retail stores. For example, Proctor and Gamble, a large household-products firm, sells all kinds of cleaning products, not just one or two. Sometimes these products are complements (e.g., laundry detergent, bleach, and fabric softeners), whereas other products are specialized substitutes.

If cost data are known, a quantitative measure of economies of scope can be determined. Consider a firm that can produce both stationary and notebook paper. The cost is \$50,000 per 1,000 reams of stationary and \$30,000 per 1,000 reams of notebook paper if the firm produces only one of these products. However, 1,000 reams of each type of paper can be produced for a total of \$70,000 if both are produced together.

A measure of economies of scope (S) is

$$S = \frac{TC(Q_A) + TC(Q_B) - TC(Q_A, Q_B)}{TC(Q_A, Q_B)}$$

where $TC(Q_A)$ is the cost of producing product A alone, $TC(Q_B)$ is the total cost of producing Q_B units of product B alone, and $TC(Q_A, Q_B)$ is the total cost of producing both A and B. Given the data on the paper firm, the extent of economies of scope is

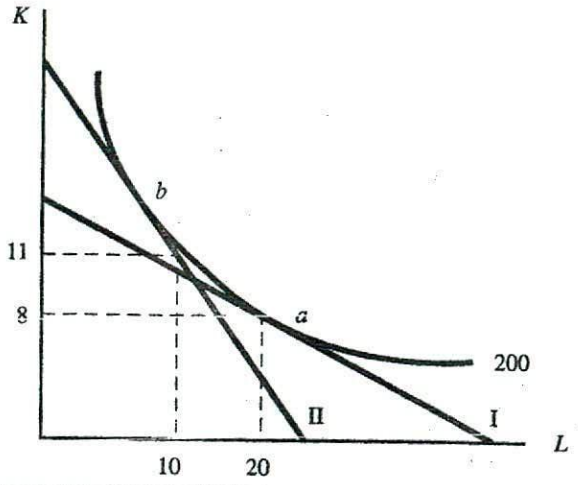
$$S = \frac{50{,}000 + 30{,}000 - 70{,}000}{70{,}000} = 0.14$$

or a 14 percent reduction in total costs associated with producing both products instead of just one. Clearly, a firm that can take advantage of economies of scope can have lower costs than other firms. In a competitive market, aggressive decision makers always will be looking for ways to capture such economies.

Factor Productivity

In most developed countries, labor quality, as measured by education and training, has been increasing at the same time that the quality of capital goods has been rising. However, because the production process involves two or more inputs working jointly to create output, it can be difficult to measure changes in the productivity of any one factor over time.

A common mistake is to simply divide output by input at two points in time and to ascribe the difference in the ratio as an increase in the productivity of that input. For example, in one production period, a firm produced 200 units of output using 10 units of labor (or 20 units of output per labor unit) and, in a later period, 240 units of output were produced using 11 units of labor, or 21.8 units of output per labor unit. One might



argue that labor productivity, as measured by output per unit of labor, had increased by 9 percent with no consideration given to the amount of capital used and/or potential changes in the productivity of that capital.

The problem is demonstrated in Figure 6.12, where in period 1 the firm is producing 200 units of output at point a using 20 units of labor and 8 units of capital. Because of a change in relative prices, the isocost function shifts from I to II, and the firm continues to produce an output rate of 200 but with a new input mix, $L = 10$ and $K = 11$. Capital has been substituted for labor due to the change in relative prices. Now, based on the ratio of output to labor (which has gone from 10 to 20), one might argue that labor productivity has doubled, but, in reality, labor quality and its true productivity is unchanged.

The use of a total factor productivity approach provides a meaningful measure of the joint productivity of the inputs. Total factor productivity is given by

$$\rho = \frac{Q}{rK + wL}$$

where r and w are input prices in the base year. Given the following data for two periods,

$Q_1 = 500$	$K_1 = 20$	$L_1 = 40$	$r_1 = 2$	$w_1 = 4$
$Q_2 = 600$	$K_1 = 22$	$L_1 = 43$		

it follows that

$$\rho_1 = \frac{500}{(2)(20) + (4)(40)} = 2.50$$

and

$$\rho_2 = \frac{600}{(2)(22) + (4)(43)} = 2.78$$

From period 1 to 2, output per dollar of input increased from 2.50 to 2.78, thus, there was an 11.2 percent increase in total factor productivity.

Key Concepts

- *Returns to scale* refers to the change in output relative to proportionate changes in inputs.
- Returns to scale are said to be: Increasing if output increases more than in proportion to the change in inputs. Decreasing if output increases less than in proportion to the change in inputs. Constant if the change in output is proportionate to the change in inputs.
- *Economies of scope* refers to per-unit cost reductions that occur when a firm produces two or more products instead of just one.
- Factor productivity refers to the ratio of output to the value of all inputs.

ESTIMATING THE PRODUCTION FUNCTION

The principles of production theory just developed are fundamental to an understanding of economics and provide an important conceptual framework for analyzing managerial problems. However, short-run output decisions and long-run planning often require more than just this conceptual framework. That is, quantitative estimates of the parameters of the production function are required for some decisions.

The general approach to estimating production functions is the same as that for most estimation problems. One of the first tasks is to select a functional form, that is, the specific relationship among the relevant economic variables. Although a variety of functional forms have been used to describe production relationships, only the Cobb–Douglas production function is discussed here. Recall that the general Cobb–Douglas function is of the form

$$Q = AK^{\alpha}L^{\beta}$$

where A, α, and β are the parameters that, when estimated, describe the quantitative relationship between the inputs (K and L) and output (Q).

The marginal products of capital and labor are functions of the parameters A, α, and β and the rates of the capital and labor inputs. That is,

$$MP_K = \frac{\partial Q}{\partial K} = \alpha AK^{\alpha-1}L^{\beta}$$

$$MP_L = \frac{\partial Q}{\partial L} = \beta AK^{\alpha}L^{\beta-1}$$

It was shown earlier that the sum of these parameters $(\alpha + \beta)$ can be used to determine returns to scale. That is,

$(\alpha + \beta) > 1 \Rightarrow$ increasing returns to scale,
$(\alpha + \beta) = 1 \Rightarrow$ constant returns to scale, and
$(\alpha + \beta) < 1 \Rightarrow$ decreasing returns to scale.

Having numerical estimates for the parameters of the production function provides significant information about the production system under study. The marginal products for each input and returns to scale can all be determined from the estimated function.

The Cobb–Douglas function does not lend itself directly to estimation by the regression methods described in chapter 4 because it is a nonlinear relationship. Technically, an equation must be a linear function of the parameters in order to use the ordinary least-squares regression method of estimation. However, a linear equation can be derived by taking the logarithm of each term. That is,

$$\log Q = \log A + a \log K + b \log L \tag{6-24}$$

That this is simply a linear relationship can be seen by setting

$$y = \log Q, \qquad A^* = \log A, \qquad X_1 = \log K \qquad X_2 = \log L$$

and rewriting the function as

$$y = A^* + aX_1 + bX_2 \tag{6-25}$$

This function can be estimated directly by the least-squares regression technique, and the estimated parameters can be used to determine all the important production relationships. Then the antilogarithm of both sides can be taken, which transforms the estimated function back to its conventional multiplicative form, as demonstrated in the following example.

Example Estimating a Production Function

Given the following data on output and inputs for 10 production periods:

Production Period	*Output (Q)*	*Capital (K)*	*Labor (L)*
1	225	10	20
2	240	12	22
3	278	10	26
4	212	14	18
5	199	12	16
6	297	16	24
7	242	16	20
8	155	10	14
9	215	8	20
10	160	8	14

1. Estimate the parameters (A, α, and β) of a Cobb–Douglas production function using the least-squares regression method.
2. Use the estimated parameters to determine
   a. Returns to scale
   b. Equations for the marginal product of labor and capital
3. Calculate the marginal products of capital and labor for the input combination $\{K = 20, L = 30\}$.

Solution

1. First, transform the production function by taking the natural logarithm of each term in the function, that is,

$$\ln Q = \ln A + \alpha \ln K + \beta \ln L$$

By transforming the output, capital, and labor data into logarithms, the least-squares regression method can be used to estimate the parameters. Using a standard multiple regression computer program, the following results were obtained:

$$\ln Q = 2.322 + 0.194 \ln K + 0.878 \ln L \qquad R^2 = 0.97$$

The estimated parameters are $\alpha = 0.194$, $\beta = 0.878$, and the value of A is determined by taking the antilogarithm of 2.322, which is 10.2. Thus the estimated production function in its original functional form is

$$Q = 10.2K^{0.194}L^{0.878}$$

2. Returns to scale are increasing because $\alpha + \beta = 1.072$ is greater than 1. The marginal product functions for capital and labor are

$$MP_K = \alpha AK^{\alpha-1}L^{\beta} = 0.194(10.2)K^{-0.806}L^{0.878}$$

and

$$MP_L = \beta AK^{\alpha}L^{\beta-1} = 0.878(10.2)K^{0.194}L^{0.122}$$

3. Substituting the estimated values for the parameters A, α, and β and the specified values of capital and labor ($K = 20$ and $L = 30$) yields the following marginal products:

$$MP_K = 0.194(10.2)(20)^{-0.806}(30)^{0.878} = 3.50$$
$$MP_L = 0.878(10.2)(20)^{0.194}(30)^{-0.122} = 10.58$$

These estimates mean that a one-unit change in capital with labor held constant at 30 would result in a 3.50-unit change in output, and a 1-unit change in labor with capital held constant at 20 would be associated with a 10.58-unit change in output.

Case Study

Empirical Estimates of Production Functions

There are many empirical studies of production functions in the United States and in other countries. One comprehensive study of a number of manufacturing industries was made by John R. Moroney. He estimated the production function

$$Q = AK^{\alpha}L_1^{\beta}L_2^{\delta}$$

where K is the dollar value of capital, L_1 is production worker-hours, and L_2 is nonproduction worker-hours. The data were taken from the Census of Manufactures, a

Estimated Production Elasticities for 17 Industries

Industry	*Estimate of* α	β	δ	α + β + δ	R^2
Food and beverages	0.555*	0.438*	0.076*	1.070*	0.987
Textiles	0.121	0.549*	0.335*	1.004	0.991
Apparel	0.128	0.437*	0.477*	1.041*	0.982
Lumber	0.392*	0.504*	0.145	0.041	0.951
Furniture	0.205	0.802*	0.103	1.109*	0.966
Paper and pulp	0.421*	0.367	0.197*	0.984	0.990
Printing	0.459*	0.045*	0.574*	1.079*	0.989
Chemicals	0.200*	0.553*	0.336*	1.090*	0.970
Petroleum	0.308*	0.546*	0.093	0.947	0.983
Rubber and plastics	0.481*	1.033*	−0.458	1.056	0.991
Leather	0.076	0.441*	0.523	1.040	0.990
Stone and clay	0.632*	0.032	0.366*	1.029	0.961
Primary metals	0.371*	0.077	0.509*	0.958	0.969
Fabricated metals	0.151*	0.512*	0.365*	1.027*	0.995
Nonelectrical machinery	0.404*	0.228	0.389*	1.020	0.980
Electical machinery	0.368*	0.429*	0.229*	1.026	0.983
Transportation equipment	0.234*	0.749*	0.041	1.023	0.972

*Indicates that the estimated parameter is significantly different from zero.

comprehensive cross-section survey of all manufacturing firms in the United States that is made every five years by the U.S. Department of Commerce.

A summary of the estimated values of the parameters of the production function (i.e., α, β, and δ) and R^2, the coefficient of determination, for each industry is shown in the following table.

Note that the R^2 values are all very high. Even the lowest, 0.951 for the lumber industry, means that more than 95 percent of the variation in output is explained by variation in the three inputs. A test of significance was made for each estimated parameter, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\delta}$, using the standard *t*-test. Those estimated production elasticities that are statistically significant at the 0.05 level are noted with an asterisk.

Of somewhat more interest is that the sum of the estimated production elasticities (α + β + δ) provides a point estimate of returns to scale in each industry. Although the sum exceeds unity in 14 of the 17 industries, it is statistically significant only in the following industries: food and beverages, apparel, furniture, printing, chemicals, and fabricated metals. Thus only in those six industries can one be confident that there are increasing returns to scale. For example, in the furniture industry, a 1 percent increase in all inputs is estimated to result in a 1.109 percent increase in output. ■

SOURCE: J. R. Moroney, "Cobb–Douglas Production Functions and Returns to Scale in U.S. Manufacturing Industry," *Western Economic Journal* 6 (December, 1967): 39–51.

Key Concepts

- The Cobb–Douglas production function $Q = AK^{\alpha}L^{\beta}$ is frequently used to estimate production relationships.
- To estimate a Cobb–Douglas production function, the function must be transformed into a linear relationship by taking the logarithm of each term (i.e., $\log Q = \log A + \alpha \log K + \beta \log L$) and then the regression estimation procedure is applied.

SUMMARY

Firms buy capital and labor and transform them through the production process into outputs of goods and services to meet the demands of consumers and other firms. The production function is an engineering or technological concept that specifies the maximum rate of output obtainable with given rates of input of capital and labor. Decisions on optimal production rates and/or efficient input combinations require that data on input prices be combined with the information generated by the production function.

The short run is defined as the period during which the rate of input use of one factor of production is fixed. By varying the rate of input of the other factor, total, average, and marginal product functions are determined. The law of diminishing marginal returns states that as more of a variable input is combined with a fixed input, a point will be reached where marginal product declines. The marginal revenue product function (*MRP*), found by multiplying marginal product by marginal revenue, is the firm's demand curve for an input. The profit-maximizing firm will hire an input until *MRP* equals the price of the input.

An isoquant is derived from the production function and shows all combinations of labor and capital that will produce a given rate of output. The isocost line shows all combinations of capital and labor that can be purchased for a given cost. If one input price changes, the slope of the isocost line will change.

The firm faces one of three production problems: (1) maximize output subject to a cost constraint, (2) minimize cost subject to an output constraint, or (3) produce that output rate that will maximize profit. Regardless of the problem addressed, the optimal input combination is determined by the tangency of an isoquant and isocost curve. At that point, the slope of the isoquant [the marginal rate of technical substitution (*MRTS*)] equals the ratio of the input prices. The firm's expansion path is defined as those points that satisfy the tangency condition.

For production to be efficient, the marginal product per dollar of input cost must be the same for all inputs. If that condition is not met, there is some way to substitute one input for the other and increase output at no additional cost. In addition, profit maximization requires that all inputs be hired until the marginal revenue product of each input equals the input's price.

The concept of returns to scale refers to the change in output associated with proportionate changes in all inputs. Such returns are increasing, decreasing, or constant, depending on whether output increases more than in proportion, less than in proportion,

or in proportion to the input changes. Increasing returns to scale may be explained by production technology that is cost-efficient only at high output rates, specialization of labor, and by inventory economies. Decreasing returns to scale may occur when firms grow so large that they are difficult to manage. Economies of scope refer to a reduction in per-unit costs when a firm increases the number of products that it produces. Factor productivity is measured by the ratio of output to the value of inputs.

Decision making often requires a quantitative estimate of the parameters of the production function. Having quantitative estimates of the parameters of a production function allows determination of the marginal product of each input and economies of scale. For a Cobb–Douglas production function, returns to scale are constant, increasing, or decreasing, depending on whether the sum of the estimated exponents is equal to 1, greater than 1, or less than 1, respectively.

Discussion Questions

6-1. Explain the concept of a production function. Why is only having qualitative information about the production function inadequate for making decisions about efficient input combinations and the profit-maximizing rate of output?

6-2. Explain the law of diminishing marginal returns and provide an example of this phenomenon.

6-3. What is the difference between the short run and the long run? What are examples of a firm where the short run would be quite short (e.g., a few days or weeks) and where it would be quite long (e.g., several months or a year or more)? Explain.

6-4. What is meant by the statement that "firms operate in the short run and plan in the long run?" Relate this statement to the operation of the college or university that you are attending.

6-5. Legislation in the United States requires that most firms pay workers at least a specified minimum wage per hour. Use principles of marginal productivity to explain how such laws might affect the quantity of labor employed.

6-6. What would the isoquants look like if all inputs were nearly perfect substitutes in a production process? What if there was near-zero substitutability between inputs?

6-7. Explain why the isocost function will shift in a parallel fashion if the cost level changes, but the isocost will pivot about one of the intercepts if the price of either input changes.

6-8. Suppose wage rates at a firm are raised 10 percent. Use theoretical principles of production to show how the relative substitution of one input for another occurs as a result of the increased price of labor. Provide an example of how input substitution has been made in higher education.

6-9. When estimating production functions, what would be some of the problems of measuring output and inputs for each of the following?
   a. A multiproduct firm
   b. A construction company
   c. An entire economy

Problems

6-1. Use the production function

$$Q = 10K^{0.5}L^{0.6}$$

to complete the following production table.

Rate of Capital Input (K)

6	24.5			56.3		71.8
5						
4		30.3				
3					45.5	
2			27.3			
1	10.0					29.3
	1	2	3	4	5	6

Rate of Labor Input (L)

a. For this production system, are returns to scale decreasing, constant, or increasing? Explain.

b. Suppose the wage rate is \$28, the price of capital also is \$28 per unit, and the firm currently is producing 30.3 units of output per period using four units of capital and two units of labor. Is this an efficient resource combination? Explain. What would be a more efficient (not necessarily the best) combination? Why? (*HINT:* Compare the marginal products of capital and labor at the initial input combination.)

6-2. Use the data from problem 6-1 to answer the following questions.

a. If the rate of capital input is fixed at three and if output sells for \$5 per unit, determine the total, average, and marginal product functions and the marginal revenue product function for labor in the following table.

L	TP_L	AP_L	MP_L	MRP_L
0				
1				
2				
3				
4				
5				
6				

b. Using the data from part (a), if the wage rate is \$28 per unit, how much labor should be employed?

c. If the rate of labor input is fixed at 5 and the price of output is \$5 per unit, determine the total, average, and marginal product functions for capital and the marginal revenue product of capital in the following table.

K	TP_K	AP_K	MP_K	MRP_K
0				
1				
2				
3				
4				
5				
6				

d. Using the data from part (c), if the price of capital is \$40 per unit, how many units of capital should be employed?

6-3. International Publishing has kept the following data on labor input and production of textbooks for each of eight production periods.

Production Period	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Labor Input	4	3	6	8	2	7	5	1
Output of Books (total product)	260	190	310	240	110	290	300	50

a. Use the data on labor input and total product to compute the average and marginal product for labor input rates from one to eight. (Assume that a zero labor input would result in zero output.)

b. Using two graphs similar to that of Figure 6.3, plot the total product function in the upper graph and average and marginal product functions in the lower graph. On the graph identify the rate of labor input: (1) where total output is at a maximum and the corresponding point where marginal product is zero; (2) where there is an inflection point on the total product function and the corresponding point where the marginal product function is at a maximum; and (3) where the slope of a line drawn through the origin to a point on the total product function would have maximum slope and the corresponding point where the average product curve is at a maximum.

c. Given that the objective of the firm is to maximize profit, can you determine from these data how much output should be produced? If not, what additional information would you need? Can you think of any circumstance where the firm would use more than six units of labor per period in this production process?

6-4. The marginal product of labor function for International Trading Inc. is given by the equation

$$MP_L = 10\frac{K^{0.5}}{L^{0.5}}$$

Currently, the firm is using 100 units of capital and 121 units of labor. Given the very specialized nature of the capital equipment, it takes six to nine months to increase the capital stock, but the rate of labor input can be varied daily. If the price of labor is \$10 per unit and the price of output is \$2 per unit, is the firm operating efficiently in the short run? If not, explain why, and determine the optimal rate of labor input.

6-5. For each of the following production functions, determine whether returns to scale are decreasing, constant, or increasing.

a. $Q = 2K + 3L + KL$
b. $Q = 20K^{0.6}L^{0.5}$
c. $Q = 100 + 3K + 2L$
d. $Q = 5K^{a}L^{b}$, where $a + b = 1$
e. $Q = 10K^{a}L^{b}$, where $a + b = 1.2$
f. $Q = K/L$

6-6. The revenue department of a state government employs certified public accountants (CPAs) to audit corporate tax returns and bookkeepers to audit individual returns. CPAs are paid $31,200 per year, while the annual salary of a bookkeeper is $18,200. Given the current staff of CPAs and bookkeepers, a study made by the department's economist shows that adding one year of a CPA's time to audit corporate returns results in an additional tax collection of $52,000. In contrast, an additional bookkeeper adds $41,600 per year in additional tax revenue.

a. If the department's objective is to maximize tax revenue collected, is the present mix of CPAs and bookkeepers optimal? Explain.

b. If the present mix of CPAs and bookkeepers is not optimal, explain what reallocation should be made. That is, should the department hire more CPAs and fewer bookkeepers or vice versa?

6-7. The production function for Superlite Sailboats, Inc., is

$$Q = 20K^{0.5}L^{0.5}$$

with marginal product functions

$$MP_K = 10\frac{L^{0.5}}{K^{0.5}} \quad \text{and} \quad MP_L = 10\frac{K^{0.5}}{L^{0.5}}.$$

a. If the price of capital is $5 per unit and the price of labor is $4 per unit, determine the expansion path for the firm.

b. The firm currently is producing 200 units of output per period using input rates of $L = 4$ and $K = 25$. Is this an efficient input combination? Why or why not? If not, determine the efficient input combination for producing an output rate of 200. What is the capital-labor ratio?

c. If the price of labor increases from $4 to $8 per unit, determine the efficient input combination for an output rate of 200. What is the capital-labor ratio now? What input substitution has the firm made?

6-8. For the production function $Q = 20K^{0.5}L^{0.5}$ determine four combinations of capital and labor that will produce 100 and 200 units of output. Plot these points on a graph and use them to sketch the 100- and 200-unit isoquants.

6-9. Suppose the price of one unit of labor is $10 and the price of a unit of capital is $2.50.

a. Use this information to determine the isocost equations corresponding to a total cost of $200 and $500.

b. Plot these two isocost lines on a graph.

c. If the price of labor falls from $10 per unit to $8 per unit, determine the new $500 isocost line and plot it on the same diagram used in part (b).

6-10. Consider two firms, A and B, with the following production functions:
Firm A: $Q_A = 100K^{0.8}L^{0.2}$ Firm B: $Q_B = 100K^{0.5}L^{0.5}$

a. If both firms use 25 units of capital and 25 units of labor, what is the output rate for each firm?

b. If the input prices are $r = \$1$ and $w = \$1$, is the input combination $K = 25$ and $L = 25$ efficient for firm A? For firm B?

c. If the input combination $K = 25$ and $L = 25$ is not efficient for either firm A or firm B, determine the efficient ratio of the two inputs for each firm. (*NOTE:*

Do not calculate unique values for K and L but only the ratio of K to L that efficient; that is, find the input ratio defined by the expansion path.)

6-11. The Economic Planning Department at International Chemicals, Inc. has used r gression analysis to estimate the firm's production function as

$$\ln Q = 3 + 0.25 \ln K + 0.75 \ln L$$

where "ln" denotes the natural logarithm of the variable.

a. Convert this production back to its original (i.e., multiplicative or Col Douglas) form. (*HINT:* Find the antilogarithm of both sides of the equatic
b. If the capital stock is fixed at 16, the price of labor is $200 per unit, and price of the firm's only product, sulfuric acid, is $10 per unit, determine the of labor input that will maximize the firm's profit.

6-12. In one production period, a firm produced an output rate of 1,000 using 50 u of capital and 40 units of labor. In a later period, output was 1,500 units, the ital input was 60 units, and the labor input was 45 units. The base period i prices are $r = 5$ and $w = 10$. Determine total factor productivity in each pe and the percentage of change in that productivity between the two periods.

6-13. A firm has determined that it can produce 100 units of output with any of the lowing input combinations:

Capital	*Labor*
20	1
16	2
12	3
11	4
9	6
7	9
5	13

a. What is the marginal rate of technical substitution between 3 and 4 units of bor? What is it between 5 and 7 units of capital?
b. Can the marginal product of labor be determined from this data? Explain.
c. Assuming there are constant returns to scale, what output rate will be pr duced if capital is 24 and labor is 6?

Problems Requiring Calculus

6-14. Given the production function

$$Q = AK^{\alpha}L^{\beta}N^{\delta}$$

where Q is the rate of output and K, L, and N represent inputs of capital, labo and land, respectively, determine

a. The specific conditions (i.e., values of α, β, and δ) under which returns to scal would be increasing, constant, and decreasing.
b. The equation for the marginal product function for each input.

6-15. A production process uses only one input, labor, and is described by the follow ing production function:

$$Q = 25L^2 - \frac{L^3}{3}$$

(*NOTE:* This function is applicable only for labor input rates between 0 and 75.) Over what output ranges are marginal returns increasing, decreasing (but still positive), and negative?

6-16. Squaretire, Inc., a small producer of automobile tires, has the following production function:

$$Q = 100K^{0.5}L^{0.5}$$

During the last production period, the firm operated efficiently and used input rates of 100 and 25 for capital and labor, respectively.

a. What is the marginal product of capital and the marginal product of labor based on the input rates specified?
b. If the price of capital was $20 per unit, what was the wage rate?
c. For the next production period, the price per unit of capital is expected to increase to $25, while the wage rate and the labor input will remain unchanged under the terms of the labor contract with the United Rubber Workers Local No. 25. If the firm maintains efficient production, what input rate of capital will be used?

6-17. Given the production function

$$Q = 30K^{0.7}L^{0.5}$$

and input prices $r = 20$ and $w = 30$,

a. Determine an equation for the expansion path.
b. What is the efficient input combination for an output rate of $Q = 200$? For $Q = 500$?

6-18. The production function for International Foodstuffs is

$$Q = 20K^{0.5}L^{0.5}$$

The initial prices of the inputs are $w = 20$ and $r = 30$. Under the labor contract with a national union, at least the current employment level of 300 workers must be maintained through the next production period. (However, more workers can be hired if necessary.)

a. In the previous production period, the firm produced 4,899 units of output. Assuming efficient production, what was the rate of capital input?
b. Because of the national recession, the desired level of output for the next production period is only 4,000 units. What is the optimal rate of capital input?

6-19. Output for the Cloverdale Farm in southern Iowa is given by the production function

$$Q = 10L^{0.7}N^{0.3}$$

where Q is the rate of output, and L and N are the rates of labor and land, respectively. In the previous production period, the farm produced 2,633 units of output using 500 acres of land (the entire cropland available) and 200 units of labor. As the result of the farm's participation in a government soil-bank program, 100 acres must be left fallow (i.e., unused) in the next production period.

a. What will be the rate of output in the next period if the same labor input rate is used together with the reduced land input?
b. How much additional labor would be needed to maintain output at the rate achieved in the original period? Explain.

6-20. The following table shows the relationship between hours of study and final examination grades in each of three classes for a particular student, who has a total of 15 hours to prepare for these tests. If the objective is to maximize the average grade in the three classes, how many hours should this student allocate to preparation for each of these classes? Explain your approach to this problem.

Managerial Economics		*History*		*Chemistry*	
Hours	*Grade*	*Hours*	*Grade*	*Hours*	*Grade*
0	40	0	50	0	30
1	50	1	60	1	50
2	59	2	69	2	60
3	67	3	77	3	66
4	74	4	84	4	71
5	79	5	90	5	74
6	83	6	95	6	76
7	86	7	96	7	77
8	88	8	97	8	77
9	89	9	97	9	77
10	89	10	97	10	77

6-21. Worldwide Fabricating manufactures metal office furniture with the following production function:

$$Q = 20K^{0.1}L^{0.9}$$

The firm currently is producing efficiently using 20 units of capital and 50 units of labor.
a. What is the rate of output?
b. What are the relative prices of capital and labor (i.e., what is the ratio of the two input prices?) Can you determine the actual price of labor and capital? Explain.
c. If output sells for $200 per unit, can you determine the firm's profit? Why or why not?

Computer Problems

The following problems can be solved by using the TOOLS program (downloadable from www.prenhall.com/petersen) or by using other computer software.

6-22. The capital and labor necessary to produce various quantities of bicycles are shown below.

Production Period	*Quantity*	*Labor*	*Capital*
1	1,100	65	40
2	660	35	15
3	1,200	75	45
4	1,000	60	30
5	900	55	30
6	840	45	25
7	1,050	60	35
8	500	30	10
9	1,130	65	45
10	700	40	15

a. Use regression analysis to estimate quantity as a multiplicative (i.e., Cobb–Douglas) function of labor and capital. Determine the estimated equation, *t*-statistics, and the coefficient of determination. What does the estimated equation imply about returns to scale? What are the marginal product equations?
b. Let the cost of capital be \$15 and the wage rate be \$20. Determine the equation for the expansion path. How much labor and capital should a firm use to produce 200 units of output efficiently?

6-23. Following are data on gross national product, labor input, and capital for the Taiwanese manufacturing sector for 1958–1972.

Year	*GNP*	*Labor*	*Capital*
1958	8,911.4	281.5	120,753
1959	10,873.2	284.4	122,242
1960	11,132.5	289.0	125,263
1961	12,086.5	375.8	128,539
1962	12,767.5	375.2	131,427
1963	16,347.1	402.5	134,267
1964	19,542.7	478.0	139,038
1965	21,075.9	553.4	146,450
1966	23,052.0	616.7	153,714
1967	26,182.2	695.7	164,783
1968	29,563.7	790.3	176,864
1969	33,376.6	816.0	188,146
1970	38,354.3	848.4	205,841
1971	46,868.3	873.1	221,748
1972	54,308.0	999.2	239,715

a. Use regression analysis to estimate a multiplicative (i.e., Cobb–Douglas) production function for Taiwan. Which coefficients are significant at the 0.05 level? What proportion of the variation in GNP is explained by variation in capital and labor?
b. Does the production process exhibit constant, increasing, or decreasing returns to scale? Explain.

CHAPTER

7 Cost Theory and Analysis

- **Preview**
- **The Economic Concept of Cost**
  - Opportunity Costs
  - Explicit and Implicit Costs
  - Normal Profit and Costs
  - Marginal, Incremental, and Sunk Costs
  - The Cost of Long-Lived Assets
- **Production and Cost**
- **Short-Run Cost Functions**
- **Long-Run Cost Functions**
- **Special Topics in Cost Theory**
  - Profit Contribution Analysis
  - Operating Leverage
- **Estimating Cost Functions**
  - Short-Run Cost Functions
  - Long-Run Cost Functions
- **Summary**
- **Discussion Questions**
- **Problems**

PREVIEW

The theory of cost, together with the principles of demand and production, constitute three of the basic areas of managerial economics. Few significant resource-allocation decisions are made without a thorough analysis of costs. For the profit maximizing firm, the decision to add a new product is made by comparing additional revenues to the additional costs associated with that new product. Similarly, decisions on capital investment (e.g., new machinery or a warehouse) are made by comparing the rate of return on the investment with the opportunity cost of the funds used to make the capital acquisitions. Costs are also important in the nonprofit sector. For example, to obtain funding for a new dam, a government agency must demonstrate that the value of the benefits of the dam, such as flood control and water supply, exceeds the cost of the project.

This chapter focuses on those principles of cost theory integral to decisions about optimal price and output rates. In contrast to the traditional approach to costs where historic cost data are typically used, the economist focuses on the concept of opportunity cost. In the first section, this economic concept of cost is developed. Next, the link between production theory and the principles of cost is developed. It is shown that efficient resource combinations for producing specific rates of output can be translated into cost data. Cost functions are then developed for both the short-run and long-run cases. Then, two special cost-related topics are discussed—profit contribution analysis and the principle of operating leverage. Finally, the methods used to empirically estimate cost functions are developed and applied.

THE ECONOMIC CONCEPT OF COST

Because the term *cost* has different meanings, it is essential that the term be defined precisely. As suggested previously, the traditional definition tends to focus on the explicit and historical dimension of cost. In contrast, the economic approach to cost emphasizes opportunity cost rather than historical cost and includes both explicit and implicit costs.

Opportunity Costs

Fundamental to the managerial economist is the concept of opportunity cost. The best measure of cost of a consumer product or a factor of production is what must be given up to obtain that product or factor. For example, a consumer who pays 10 dollars for dinner may have to give up going to a concert. The manager who hires an additional secretary may have to forgo hiring an additional clerk in the shipping department. Alternatively, the cost to society of adding another soldier to the army is not only the dollar outlay for salary, uniforms, and equipment but also the foregone output this individual would have produced as a civilian worker. In general, the opportunity cost of any decision is the value of the next best alternative that must be foregone.

In order to maximize the value of the firm, the effective manager must view costs from this perspective. For example, budgeting is fundamental to most organizations. The very nature of that process implies that opportunity costs are incurred whenever budget resources are allocated to one department rather than another. A reallocation from the production department to the research and development group may result in

new and better products in the future but the cost is lower production and profit for the current period. Obviously, such a decision should be made only when management is convinced that the potential for even greater profit in the future outweighs the reduced profit for the current period so that shareholder value will be enhanced.

Explicit and Implicit Costs

Sometimes, the full opportunity cost of a business decision is not accounted for because of failure to include implicit costs. In general, *explicit costs* are those costs that involve an actual payment to other parties, while *implicit costs* represent the value of foregone opportunities but do not involve an actual cash payment. Implicit costs are just as important as explicit costs but may be neglected because they are not as obvious. For example, a manager who runs his own business forgoes the salary that could have been earned working for someone else. This implicit cost generally is not reflected in accounting statements, but rational decision making requires that it be considered.

To see how reliance on historical rather than opportunity cost can lead to a poor decision, consider the following example. A bakery has an inventory of wheat that was purchased at $3 per bushel but is now worth $5 per bushel. The firm is considering using this wheat to make a new whole wheat bread that will be sold to stores for $5 per unit (six loaves). Suppose that one bushel of wheat is required to make each unit of this new type of bread, while $1.50 of labor, energy, and other costs per unit of output are also incurred.

The traditional approach to cost would value the wheat input at $3 per bushel and estimate profit on the finished product to be $0.50 per unit, as shown in Table 7.1. In contrast, the economic approach to cost would value the wheat at the current market price of $5 per bushel. Analyzing the decision to produce the new bread from this approach indicates a loss of $1.50 per unit of output. Note that the only difference in the two approaches is the value placed on the inventory of wheat.

Consider the same problem in another way. The money spent on the inventory of wheat is gone; now, how can the firm best use the inventory? That is, what will be the net revenue per bushel if the new bread is manufactured compared to selling the wheat inventory in the market without processing it? If the decision is made to manufacture the bread, the firm will have a net cash flow of $3.50 per bushel, that is, the $5 selling price less $1.50 of other costs. In contrast, if the wheat is simply sold rather than processed, the firm would receive a net cash flow of $5 per bushel. Clearly, selling the wheat rather than producing the bread is the better alternative because profits will be greater. The example demonstrates that the use of the correct cost concept is essential to sound decision making. It also suggests that costs incurred in the past generally are irrelevant when making decisions.

TABLE 7.1 Traditional versus Economic Approach to Determining Profit for Use of Owner-Supplied Inputs

	Traditional Approach	*Economic Approach*
Price of finished product	$5.00	$5.00
Less: Cost of wheat input	3.00	5.00
Less: Other costs	1.50	1.50
Net profit per unit	$0.50	−$1.50

Normal Profit and Costs

As will be demonstrated in chapter 9, in industries characterized by substantial competition, principles of economics predict that profits will be driven to zero in the long run. While this sounds inconsistent with the conventional idea that most firms report a profit each year, the apparent inconsistency disappears when the concepts of economic costs and economic profits are understood and used to measure revenues and costs.

Because all opportunity costs must be accounted for, the proper concept of cost includes a normal payment to all inputs, including managerial and entrepreneurial skills and capital supplied by the owners of the firm. A normal return to management or capital is the minimum payment necessary to keep those resources from moving to some other firm or industry. Thus, cost includes a normal rate of profit. The term *economic profit* refers to profit in excess of these normal returns. That is, economic profit is defined as revenue less all economic costs.

A firm earning zero economic profit generally would show a positive profit on the income statement prepared by its accountants. This is because the normal returns to entrepreneurial skill and capital supplied by the owners are not included as costs on that statement. Unless otherwise indicated, the term *cost* will refer to all explicit and implicit costs, and the term *profit* will refer to revenues after all economic costs, including the normal returns just described, have been subtracted.

Marginal, Incremental, and Sunk Costs

Clearly, cost is an important consideration in decision making. But as the bakery example showed, it is essential that only those costs that matter be considered. Three types of cost need to be identified: sunk costs, marginal costs, and incremental costs.

Sunk costs are expenditures that have been made in the past or that must be paid in the future as part of a contractual agreement. The cost of inventory and future rental payments on a warehouse that must be paid as part of a long-term lease are examples. In general, sunk costs are irrelevant in making decisions. For instance, suppose that the monthly rental payment on the warehouse is $1,000, but the firm finds it no longer needs the space. The firm offers to sublease the space but finds that the best offer is for $800 per month. Clearly, the firm should take that offer; the additional revenue, $800 per month, is greater than the additional cost, which is zero. The $1,000 per month payment is irrelevant because it must be made regardless of the decision to rent the warehouse space. In retrospect, the decision to enter the long-term lease was a mistake, but the costs associated with that decision are sunk and now irrelevant in the decision about what to do with the warehouse.

Marginal cost refers to the change in total cost associated with a one-unit change in output. This concept is integral to short-run decisions about profit maximizing rates of output. For example, in an automobile manufacturing plant, the marginal cost of making one additional car per production period would be the labor, materials, and energy costs directly associated with that extra car. In contrast, the term *incremental cost* refers to the total additional cost of implementing a managerial decision. The costs associated with adding a new product line, acquiring a major competitor, or developing an in-house legal staff fall into the broader class of incremental costs. In a sense, marginal cost is that subcategory of incremental cost that refers to the additional cost associated with the decision to make marginal variations in the rate of output.

In the warehouse rental example, the only incremental costs the firm faces when subleasing the property may be the cost of preparing and negotiating the details of the new rental agreement. Clearly, the $1,000 per month sunk cost is not a component of incremental cost. It is essential that incremental cost measurement be done carefully so that all possible additional costs are included, but costs that are sunk are not included.

The Cost of Long-Lived Assets

Another area where accounting and economic definitions of cost diverge is for assets such as buildings, machinery, and other types of capital equipment that may last for a number of years. These are referred to as *long-lived assets.* The traditional approach to measuring the periodic cost of these assets is to combine historical cost and one of several depreciation methods to assign part of the historical cost to each year of the defined life of the asset so that the total expenses over that life will equal the historical cost. For example, using the straight-line depreciation method, an asset costing $1,000 and having a five-year life would be depreciated at the rate of $200 per year. Thus, the total historical cost will be exhausted entirely over this five-year period. The asset may have considerable value to the firm after this period, but this is not reflected in such an accounting statement. Generally, tax guidelines and considerations dictate the decision on the depreciation method used. This approach to cost measurement is adequate if the objective is to have an arbitrary method for reporting on the flow of funds into and out of the business over some time interval or to meet income tax regulations. However, as a tool for managerial decision making, the approach is flawed.

In contrast, the economic approach determines the cost as the difference between the market value of the asset at the beginning and end of the period. If the market value of the machine just discussed was $1,000 at the beginning of the year and $600 at the end, the economic cost of using it for that period was $400, not $200 as indicated by the straight-line depreciation method. It is possible for some long-lived assets to actually increase in value over time, implying that their cost was negative for that period. This has been true for some buildings and other types of real estate.

Key Concepts

- Opportunity cost, the value of a resource in its next best use, is the best way to measure cost.
- Both explicit and implicit costs must be considered in decision making.
- Economic profit is revenue minus all costs, including normal returns to management and capital.
- In general, only incremental and marginal costs are relevant in decision making; sunk costs are of little or no importance.
- The economic cost of a long-lived asset during a period is the change in its market value from the beginning to the end of the period.

PRODUCTION AND COST

A cost function relates cost to the rate of output. The basis for a cost function is the production function and the prices of inputs. Recall from chapter 6 that the expansion path defines the efficient combination of capital and labor input rates for producing any rate of output. Thus the minimum cost of a given rate of output is found by multiplying the efficient rate of each input by their respective prices and summing the costs. The combination of that cost and the associated rate of output defines one point on the cost function.

In chapter 6 it was shown that the production function, $Q = 100K^{0.5}L^{0.5}$, has an expansion path $K = (w/r)L$. Thus if the price of labor is 2 and the price of capital is 1, the expansion path would be $K = 2L$, and the firm would expand by adding inputs at the rate of two units of capital for each additional unit of labor. Table 7.2 shows a set of efficient labor–capital combinations, the rate of output associated with each combination, and the cost of those inputs. Columns (3) and (6) represent the long-run total cost schedule for this production function, given that input prices are $w = 2$ and $r = 1$. This schedule shows the total cost of producing various rates of output. For example, the total cost of efficiently producing an output rate of 283 units per period is \$8. That is, four units of capital at a price of \$1 per unit and two units of labor at a price of \$2 per unit would cost \$8. Thus, the minimum cost of producing 424 units of output is \$12. In this example it is a long-run function because both inputs are variable—there is no fixed factor of production.

In the short run, at least one factor of production is fixed, and the cost of that input is defined as *fixed cost.* Regardless of the rate of output, that cost does not change. That is, management has made a decision in the past that obligates the firm to pay certain costs that are independent of the rate of output. A long-term lease on a warehouse, an employment contract with an executive, and a collective bargaining agreement with a labor union are examples of commitments that may result in fixed costs. Fixed costs fall into the category of sunk costs.

As suggested in chapter 6, the length of the operating period varies with the nature of the business. A small business may be able to vary all inputs in a matter of days. Conversely, adding additional capacity to a nuclear generating facility could take years. The length of this period depends on the degree of asset specialization, the economic life of the assets, the time necessary to order and install new capital equipment, and the amount of training that labor requires.

			Cost		
K	*L*	*Q*	*Capital*	*Labor*	*Total*
2	1	141	\$ 2	\$ 2	\$ 4
4	2	283	4	4	8
6	3	424	6	6	12
8	4	565	8	8	16
10	5	707	10	10	20

Asset specialization refers to the number of uses of an asset. For example, a nuclear generator can only be used to generate electricity in a nuclear power plant and is a very specialized asset. Thus it might be difficult to sell that generator because only another nuclear power plant could use it. Conversely, many trucks can be used to haul a variety of products and are rather unspecialized assets. If a firm decides that one of its trucks is no longer needed, it could easily be sold in the market and its cost eliminated.

Clearly, the longer the life of the asset, the longer the period defined as the short run. Buildings and some types of machinery may have an economic life of many years. Once in place, their costs may be fixed for a long period of time. Also, it may take months or even years to order and install certain types of capital equipment and/or to develop a particular set of skills in labor. In such cases the firm may find that part of the cost of its productive capacity is fixed for that period.

Key Concepts

- The long-run total cost of any rate of output is determined by the expansion path of the firm (which relates output rates and efficient input combinations) and the prices of the inputs.
- If one input is fixed, the costs associated with that input are called fixed costs, and the firm is said to be operating in the short run. The costs associated with the nonfixed inputs are called variable costs.

SHORT-RUN COST FUNCTIONS

Managerial decision making is facilitated by information that shows the cost of each rate of output. Consider a production process that combines variable amounts of labor with a fixed capital stock, say, 10 machines. In this process, the rate of production is changed by varying the rate of labor input. Assume that the firm can vary the labor input freely at a cost of $100 per unit of labor per period. Therefore, the expenditure for labor is the variable cost. If the 10 machines are rented under a long-term lease at $100 per machine per production period, the fixed cost would be $1,000 per period.

Table 7.3 summarizes the relevant production and cost data for this production process, and the data are shown graphically in Figure 7.1a. Note that fixed cost is indicated by a horizontal line; that is, this cost is constant with respect to output. The total variable cost function (TVC) begins at the origin, increases at a decreasing rate up to an output rate between 3 and 4, and then increases at an increasing rate. Total cost (TC) has the same shape as total variable cost but is shifted upward by $1,000, the amount of fixed cost. These functions relate an output rate to the total cost of producing that output rate.

Functions that indicate the cost per unit of output also can be determined. Often, these are more useful for decision making than are total cost functions. This is because managers must compare cost per unit of output to the market price of that output. Recall that market price is measured per unit of output. By dividing a total cost function by output, a corresponding per-unit cost function is determined. That is,

Input Rate		Rate of Output	Total Fixed Cost	Total Variable Cost	Total Cost
Capital	Labor				
10	0	0	$1,000	$ 0	$1,000
10	2.00	1	1,000	200	1,200
10	3.67	2	1,000	367	1,367
10	5.10	3	1,000	510	1,510
10	6.77	4	1,000	677	1,677
10	8.77	5	1,000	877	1,877
10	11.27	6	1,000	1,127	2,127
10	14.60	7	1,000	1,460	2,460
10	24.60	8	1,000	2,460	3,460

$$AVERAGE\ TOTAL\ COST: AC = \frac{TC}{Q} \tag{7-1}$$

$$AVERAGE\ VARIABLE\ COST: AVC = \frac{TVC}{Q} \tag{7-2}$$

$$AVERAGE\ FIXED\ COST: AFC = \frac{TFC}{Q} \tag{7-3}$$

The marginal cost per unit of output (MC) is the change in total cost associated with a one-unit change in output, that is,

$$MARGINAL\ COST: MC = \frac{\Delta TC}{\Delta Q} \tag{7-4}$$

As is true of all associated total and marginal functions, marginal cost is the slope of the total cost function.

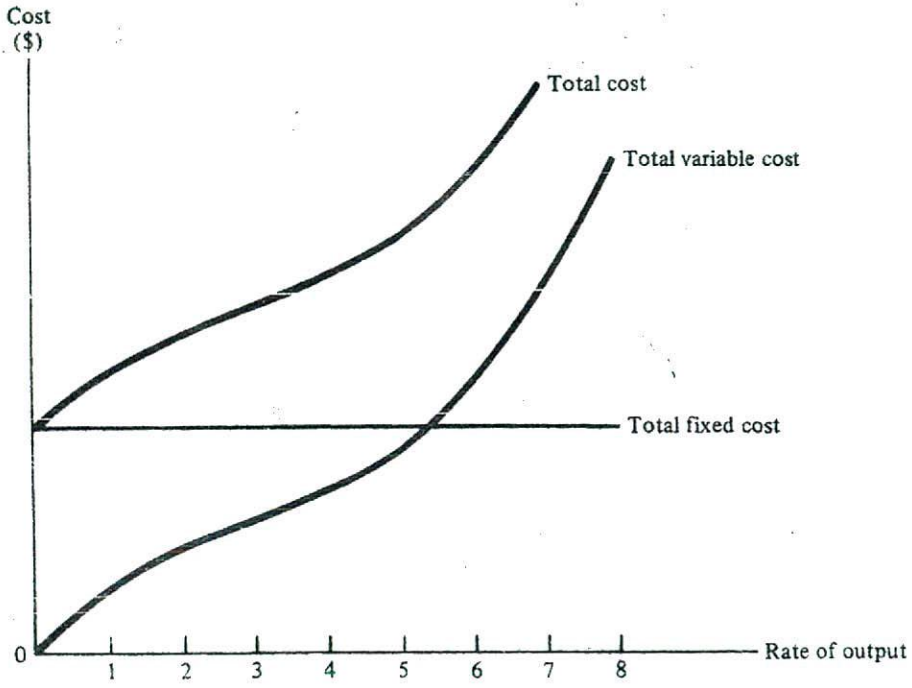
Using calculus, marginal cost is determined as the first derivative of the total cost function. That is, if the total cost function is

$$TC = f(Q)$$

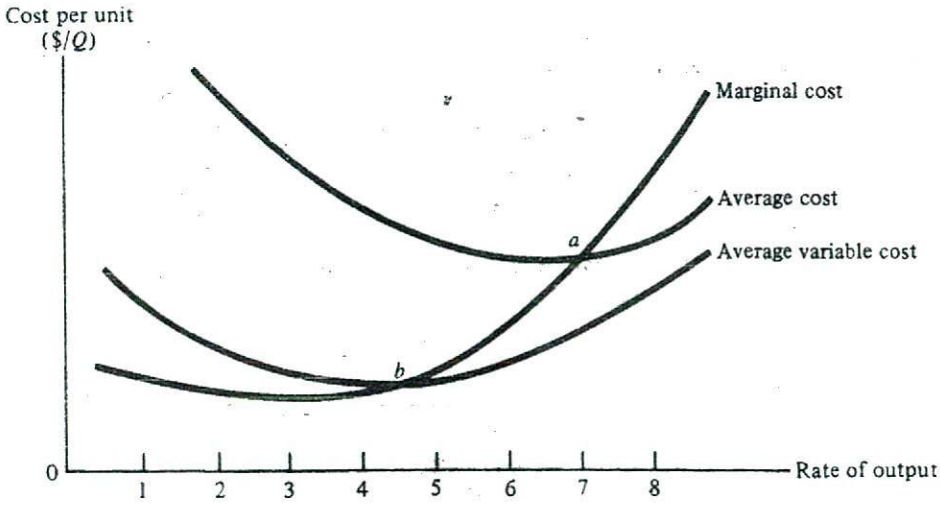
marginal cost would be

$$MC = \frac{d(TC)}{dQ} \tag{7-5}$$

Based on the total cost functions in Table 7.3, data for each per-unit cost function are reported in Table 7.4 and shown graphically in Figure 7.1b. The average total cost, average variable cost, and marginal cost functions are important in managerial decision making. In contrast, the average fixed cost function has little value for such decisions. Further, the difference between average total cost and average variable cost is average fixed cost. Thus the AC and AVC curves provide information on fixed cost per unit should such information be needed.



(a) Total cost functions



(b) Per unit cost functions

Output	*Average Fixed Cost (AFC)*	*Average Variable Cost (AVC)*	*Average Total Cost (AC)*	*Marginal Cost (MC)*
0	—	—	—	
1	$1,000	$200	$1,200	$ 200
2	500	184	684	167
3	333	170	503	143
4	250	169	419	167
5	200	175	375	200
6	167	188	355	250
7	143	209	351	333
8	125	307	432	1,000

The per-unit cost functions for many production systems have the U shape shown in Figure 7.1b. At low rates of production, there is too little of the variable input relative to the fixed input. As the variable input is increased, output rises rapidly, and the cost per unit falls. Initially, total cost increases but at a decreasing rate. This implies that marginal cost (the slope of total cost) is falling. Because of the law of diminishing marginal returns, additional units of the variable input result in smaller additions to output and ultimately marginal cost rises. When marginal cost exceeds average cost, the average cost function begins to rise.

As is true of all marginal and average functions, as long as marginal cost is below the average cost curve, the average function will decline. When marginal is above average, the average cost curve will rise. This implies that marginal cost intersects both the average total cost and average variable cost functions at the minimum point of the average curves (points *a* and *b* in Figure 7.1b).

Example Finding Minimum Average Variable Cost

Given the total cost function

$$TC = 1{,}000 + 10Q - 0.9Q^2 + 0.04Q^3$$

find the rate of output that results in minimum average variable cost.

Solution Marginal cost is the first derivative of the total cost function

$$\frac{d(TC)}{dQ} = MC = 10 - 1.8Q + 0.12Q^2$$

Now, find the total variable cost function (TVC) by subtracting the fixed cost component ($1,000) from the total cost function. That is,

$$TVC = 10Q - 0.9Q^2 + 0.04Q^3$$

Then find average variable cost (AVC) by dividing TVC by output (Q). That is,

$$AVC = \frac{TVC}{Q} = \frac{10Q - 0.9Q^2 + 0.04Q^3}{Q}$$

$$AVC = 10 - 0.9Q + 0.04Q^2$$

Because the minimum point of *AVC* occurs at its intersection with marginal cost, equate the *AVC* and *MC* functions and solve for *Q*. That is,

$$AVC = MC$$
$$10 - 0.9Q + 0.04Q^2 = 10 - 1.8Q + 0.12Q^2$$

Rearranging terms yields a quadratic equation

$$-0.08Q^2 + 0.9Q = 0$$

or

$$Q(-0.08Q + 0.9) = 0$$

which has the roots

$$Q_1 = 0 \text{ and } Q_2 = 11.25$$

Disregarding the root associated with a zero output rate, it is seen that minimum *AVC* is achieved at an output rate of 11.25 units.

Alternatively, the minimum point of *AVC* could be found by setting the first derivative of *AVC* equal to zero and solving for *Q*. That is,

$$\frac{d(AVC)}{dQ} = -0.9 + 0.08Q = 0$$
$$0.08Q = 0.9$$
$$Q = 11.25$$

Key Concepts

- Per-unit or average cost functions sometimes are more useful than total cost functions in making decisions. The average cost functions are found by dividing the relevant total cost functions by output, that is,

$$AVERAGE\ COST: AC = \frac{TC}{Q}$$

$$AVERAGE\ VARIABLE\ COST: AVC = \frac{TVC}{Q}$$

$$AVERAGE\ FIXED\ COST: AFC = \frac{TFC}{Q}$$

- Marginal cost per unit is the change in total cost associated with a one-unit change in output, that is,

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

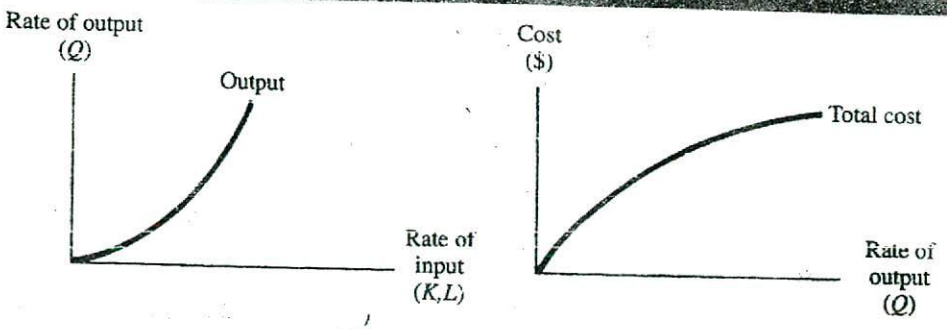
- Marginal cost intersects both the average total cost and average variable cost functions at the minimum point of the average curves.

LONG-RUN COST FUNCTIONS

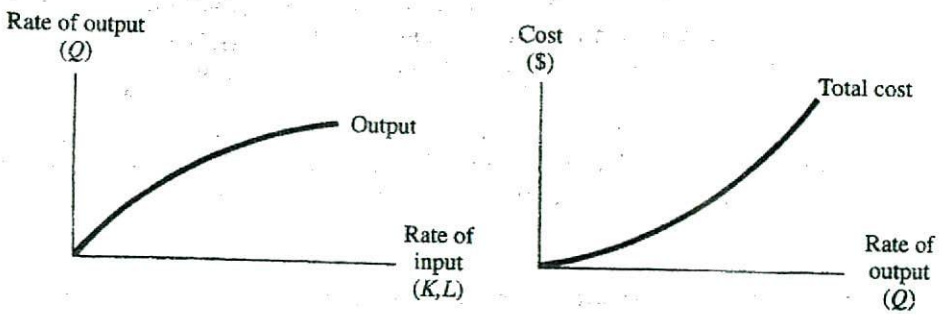
Firms operate in the short run but plan in the long run. At any point in time, the firm has one or more fixed factors of production. Therefore, production decisions must be

made based on short-run cost curves. However, most firms can change the scale of their operation in the long run by varying all inputs, and in doing so, move to a preferred short-run cost function.

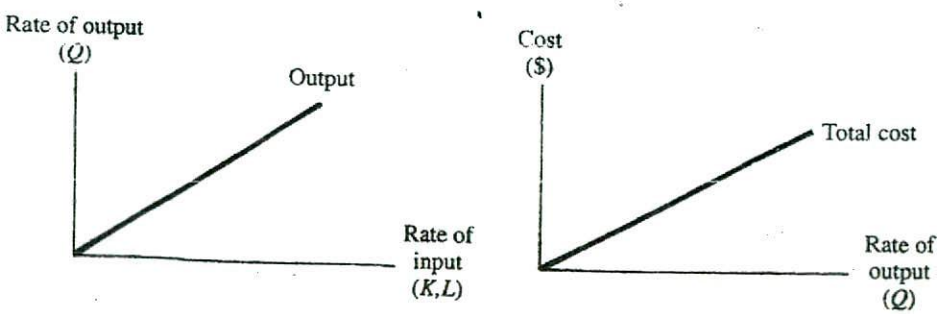
Recall from chapter 6 that returns to scale are increasing, decreasing, or constant, depending on whether a proportional change in both inputs results in output increasing more than in proportion, less than in proportion, or in proportion to the increase in inputs. These three possibilities are shown in the left-hand panels of Figure 7.2.



(a) Increasing returns to scale



(b) Decreasing returns to scale



(c) Constant returns to scale

There is a direct correspondence between returns to scale in production and the long-run cost function for the firm. If returns to scale are increasing, inputs are increasing less than in proportion to increases in output. Because input prices are constant, it follows that total cost also must be increasing less than in proportion to output. This relationship is shown in Figure 7.2a. If decreasing returns to scale apply, the total cost function increases at an increasing rate. Constant returns to scale implies that total cost will change in proportion to changes in output. The latter two relationships are shown in parts (b) and (c) of Figure 7.2.

Case Study

Economies of Scale in the Banking Industry

The past 15 years have seen numerous mergers of banks in every part of the United States. Invariably, the managers of these banks pointed to significant cost reductions (i.e., increasing returns to scale) associated with consolidation of computer systems, combining of neighboring branch outlets, and reduction of corporate overhead expenses as justification.

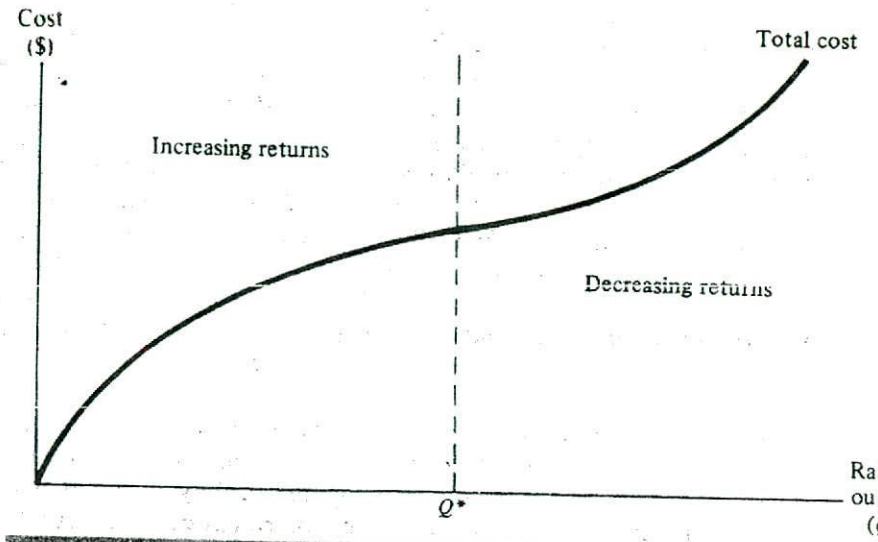
Many of these mergers involved multibillion dollar banks, which appeared to be inconsistent with existing empirical research on bank costs that showed significant diseconomies of scale for banks with more than $25–50 million in deposits.[1] Unfortunately, these studies used data only for banks with less than $1 billion in deposits. In a more recent study, Sherrill Shaffer and Edmond David used data for large banks (those with $2.5 to $121 billion in deposits) and found increasing returns to scale (i.e., declining per unit costs) up to a bank size of $15 to $37 billion.[2]

Clearly, the owners and managers of the merged banks knew more about their actual cost functions than did the earlier economic analysts. The consistent pattern of mergers of banks much larger than $24 to $50 million in deposits was strong evidence that the existing research was not correct. ■

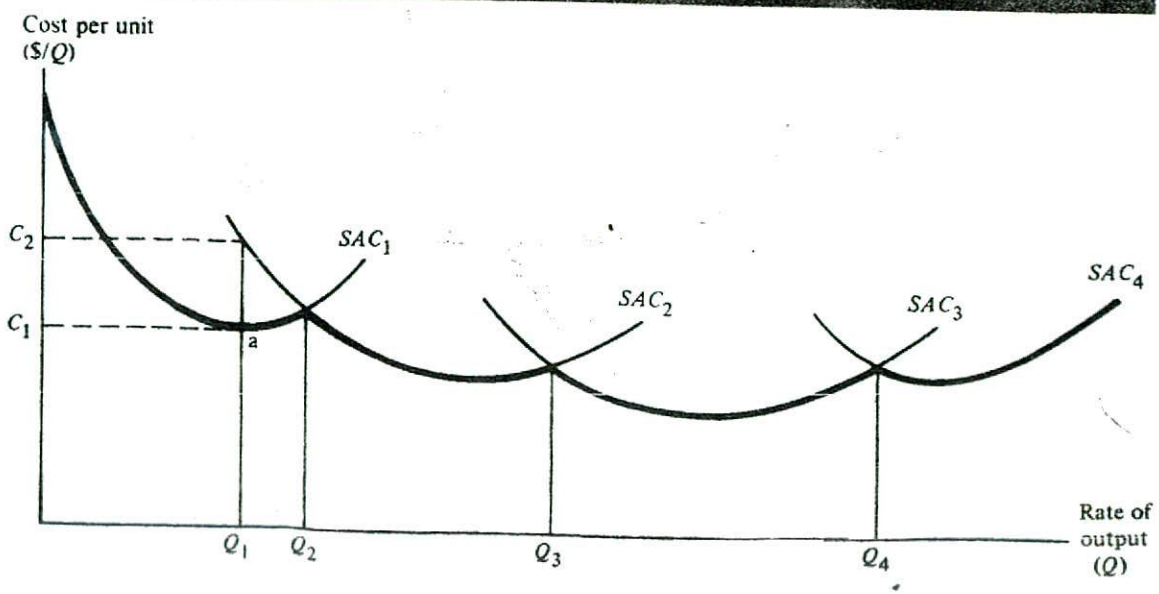
[1]See G. Benston, G. A. Hanweck, and B. Humphrey, "Scale Economies in Banking: A Restructuring and Reassessment," *Journal of Money, Credit, and Banking* 14 1982:435–56; and T. Gilligan, M. Smirlock, and W. Marshall, "Scale and Scope Economies in the Multi-Product Banking Firm," *Journal of Monetary Economics* 13 1984:393–405.

[2]S. Shaffer and E. David, "Economies of Superscale in Commercial Banking," *Applied Economics* 23 1991:283–293.

The production process of many firms is characterized first by increasing returns and then by decreasing returns. In this case, the long-run total cost function first increases at a decreasing rate and then increases at an increasing rate, as shown in Figure 7.3. Such a total cost function would be associated with a U-shaped long-run average cost function.



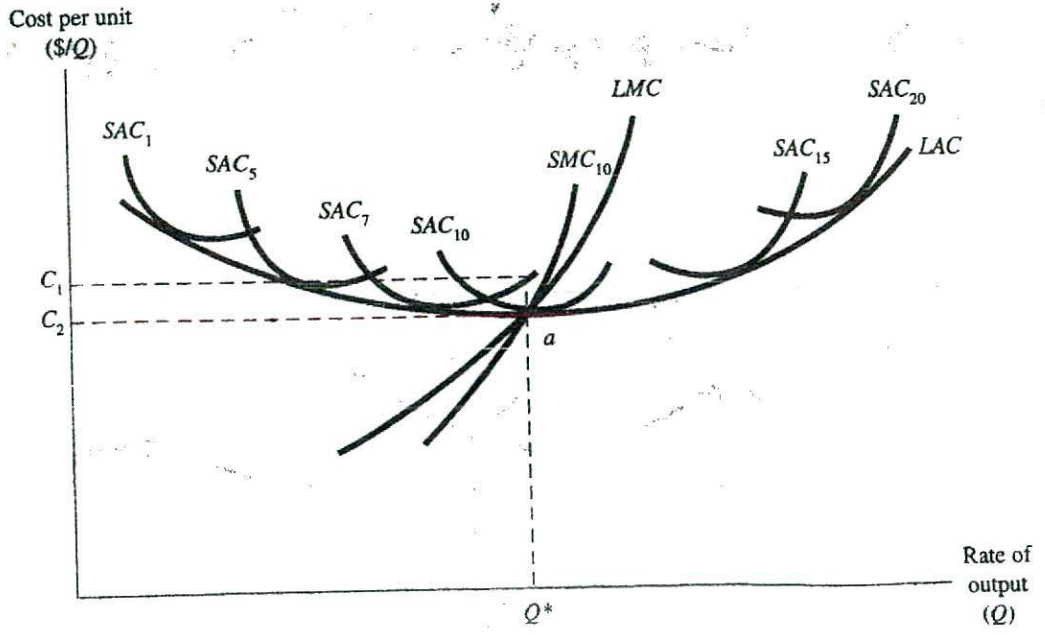
Suppose that a firm can expand the scale of operation only in discrete units. For example, the generators for large electric power plants are made in only a few sizes. Often, these power plants are built in multiples of 750 megawatts (MW). That is, output capacity of alternative plants would be 750 MW, 1500 MW, 2250 MW, and so on. The short-run average cost functions in Figure 7.4 (labeled SAC_1 through SAC_4) are associated with each of four discrete scales of operation. The long-run average cost



function for this firm is defined by the minimum average cost of each level of output. For example, output rate Q_1 could be produced by plant size 1 at an average cost of C_1 or by plant size 2 at a cost of C_2. Clearly, the cost is lower for plant size 1, and thus point *a* is one point on the long-run average cost curve. By repeating this process for various rates of output, the long-run average cost is determined. For output rates of zero to Q_2, plant 1 is the most efficient and that part of SAC_1 is part of the long-run cost function. For output rates Q_2 to Q_3, plant 2 is the most efficient, and for output rates Q_3 to Q_4, plant 3 is the most efficient. The scallop-shaped curve shown in boldface in Figure 7.4 is the long-run average cost curve for this firm. This boldfaced curve is called an *envelope curve*. Firms plan to be on this envelope curve in the long run. Consider a firm currently operating plant size 2 and producing Q_1 units at a cost of C_2 per unit. If output is expected to remain at Q_1, the firm will plan to adjust to plant size 1, thus reducing per-unit cost to C_1.

Most firms will have many alternative plant sizes to choose from, and there is a short-run average cost curve corresponding to each. A few of the short-run average cost curves for these plants are shown in Figure 7.5. Only one point or a very small arc of each short-run cost curve will lie on the long-run average cost function. Thus long-run average cost can be shown as the smooth U-shaped curve labeled *LAC*. Corresponding to this long-run average cost function is a long-run marginal cost curve *LMC*, which intersects *LAC* at its minimum point *a*, which is also the minimum point of short-run average cost curve 10. The short-run marginal cost curve (SMC_{10}) corresponding to SAC_{10} is also shown. But $SMC_{10} = SAC_{10}$ at the minimum point of SAC_{10}. Thus at point *a* and only at point *a* the following unique result occurs:

$$SAC = SMC = LAC = LMC \tag{7-6}$$



The long-run cost curve serves as a long-run planning mechanism for the firm. For example, suppose that the firm is operating on short-run average cost curve SAC_7 in Figure 7.5, and the firm is currently producing an output rate of Q^*. By using SAC_7, it is seen that the firm's cost per unit is C_1. Clearly, if projections of future demand indicate that the firm could expect to continue selling Q^* units per period, profit could be increased by increasing the scale of plant to the size associated with short-run average curve SAC_{10}. With this plant, cost per unit for an output rate of Q^* would be C_2 and the firm's profit per unit would increase by $C_1 - C_2$. Thus total profit would increase by $(C_1 - C_2) \cdot Q^*$.

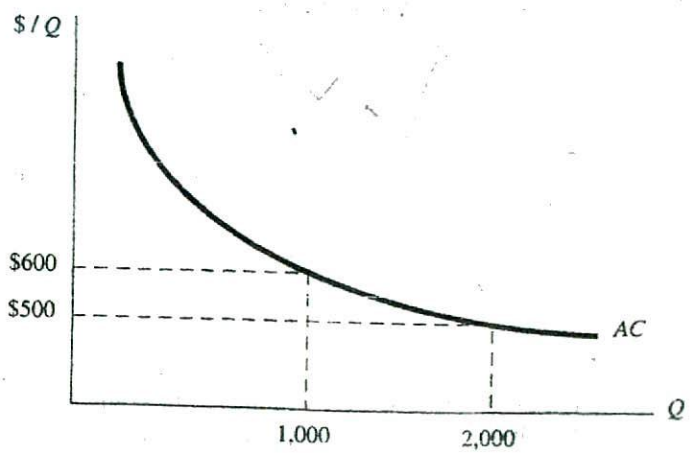
Example **Fixed Cost, Economies of Scale, and Global Market Expansion**

A firm's average total cost function is depicted in the following figure. If the firm serves only the United States market, it will sell 1,000 units, but if it also sells in Europe and Asia it can sell 2,000 units.

a. What is the explanation for the declining average cost function?
b. Assuming the firm is operating in the short run and that the average fixed cost is 20 percent of average total cost at an output rate of 1,000: (i) what is total fixed cost and what is average fixed cost at an output rate of 2,000? (ii) Is average variable cost increasing or decreasing between $Q = 1{,}000$ and $Q = 2{,}000$.

Solution

a. If the firm is operating in the long run the answer is that the production process is characterized by increasing returns to scale. If the firm is operating in the short run, part of the explanation is that fixed cost is being spread over a larger number of units produced and that marginal cost (whether increasing or decreasing) is below average total cost.
b. (i) At $Q = 1{,}000$, $AFC = \$120$ (i.e., 20 percent of \$600). Thus, total fixed cost is $Q \cdot AFC = \$1{,}000 \cdot 120 = \$120{,}000$. At $Q = 2{,}000$, $AFC = \$120{,}000/2{,}000 = \$60$. (ii) At $Q = 1{,}000$, $AC = \$600$ so total cost is $Q \cdot AC = 1{,}000 \cdot \$60 = \$600{,}000$. Subtracting total fixed cost of \$120,000 yields total variable cost of \$480,000 or an average variable cost of \$480. (ii) At $Q = 2{,}000$, total cost is \$1,000,000 (i.e., $2{,}000 \cdot 500$) and, therefore, total variable cost is \$880,000 and $AC = 440$. Thus, AVC is declining over the output range 1,000 to 2,000.



Key Concepts

- The firm's long-run average cost function will be
  Decreasing where returns to scale in production are increasing.
  Constant where returns to scale are constant.
  Increasing where returns to scale are decreasing.
- The long-run average cost function is the envelope curve consisting of points or arcs on a number of short-run average cost curves.

SPECIAL TOPICS IN COST THEORY

Cost functions are essential to making effective managerial decisions about output and prices. At this point, several extensions of cost theory that have implications for managerial decisions will be discussed. These concepts include profit contribution analysis (including the special case of breakeven analysis) and operating leverage.

Profit Contribution Analysis

The difference between price and average variable cost ($P - AVC$) is defined as *profit contribution*. That is, revenue on the sale of a unit of output after variable costs are covered represents a contribution toward profit. At low rates of output, the firm may be losing money because fixed costs have not yet been covered by the profit contribution. Thus, at these low rates of output, profit contribution is used to cover fixed costs. After fixed costs are covered, the firm will be earning a profit.

A manager may want to know the output rate necessary to cover all fixed costs and to earn a "required" profit of π_R. Assume that both price and variable cost per unit of output (AVC) are constant. Profit (π) is equal to total revenue (PQ) less the sum of total variable costs ($Q \cdot AVC$) and fixed costs. Thus

$$\pi_R = PQ - [(Q \cdot AVC) + FC]$$

Solving this equation for Q yields a relation that can be used to determine the rate of output necessary to generate a specified rate of profit. That is,

$$Q = \frac{FC + \pi_R}{P - AVC} \tag{7-7}$$

For example, suppose that $FC = \$10{,}000$, $P = \$20$, $AVC = \$15$, and that the firm has set a required profit target of $20,000. To generate this profit, an output rate of 6,000 units is required; that is,

$$Q_R = \frac{\$10{,}000 + \$20{,}000}{20 - 15} = 6{,}000$$

A special case of this equation is where the required economic profit is zero, that is, $\pi_R = 0$. This output rate is called the breakeven point for the firm. (Recall that a zero economic profit means that normal returns are being earned by capital and other factors of production.) The breakeven rate of output, Q_e, is given by the equation

$$Q_e = \frac{FC}{P - AVC} \tag{7-8}$$

Using the data just given, it is seen that the breakeven rate of output is 2,000; that is,

$$Q_e = \frac{\$10{,}000}{20 - 15} = 2{,}000$$

This example of breakeven analysis is shown graphically in Figure 7.6. Fixed cost is shown as the horizontal line at $10,000. Total cost is given by the equation

$$TC = FC + TVC$$

Because $FC = 10{,}000$ and variable cost per unit is 15, the total cost function is

$$TC = 10{,}000 + 15Q$$

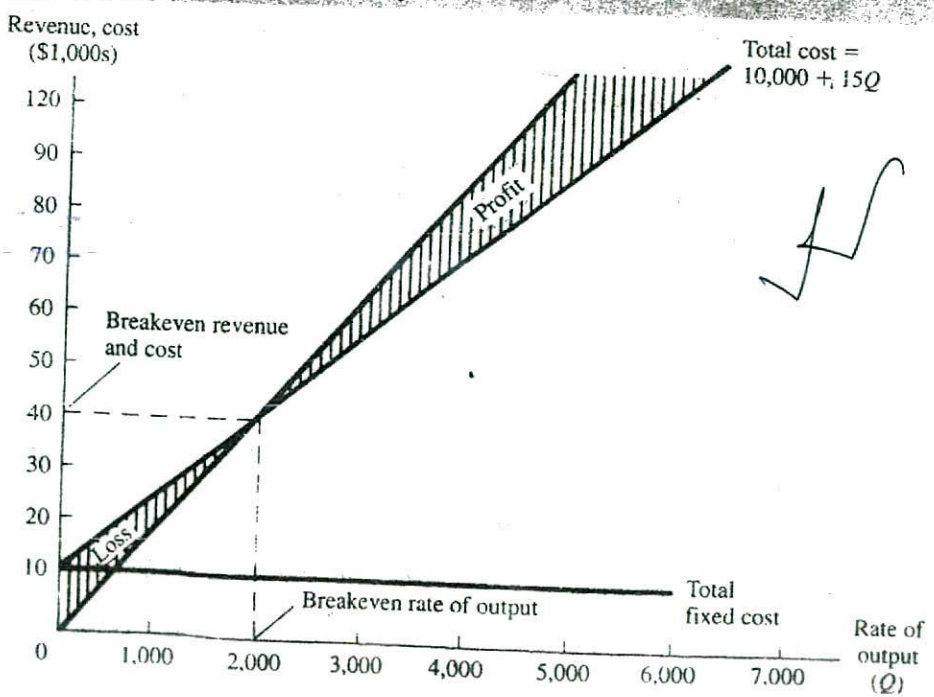
Since price is constant, the total revenue function is

$$TR = 20Q$$

which is shown as a straight line through the origin having a slope of 20. The breakeven point occurs at an output rate of 2,000, which is at the intersection of the total revenue and total cost functions. At this point, both total revenue and total cost are $40,000.

This linear approach to profit contribution has been criticized because of the assumption that both price and average variable cost are constant. The price assumption is not unrealistic for many firms, as they are able to sell all they can produce at the going

FIGURE 7.6 Linear Breakeven Analysis



price. However, other firms, especially those where sales are large relative to the size of the market, may have to reduce price to sell more output. For some firms, the assumption of constant average variable cost may be unrealistic. However, if the price and average variable cost are roughly constant, at least over the limited range of output relevant to the problem, breakeven analysis is a useful tool for managerial decisions. However, care and judgment are required in its application.

Example Breaking Even on Microcomputer Software

MicroApplications Inc. is a small firm that specializes in the production and mail-order distribution of computer programs for microcomputers. The accounting department has gathered the following data on development and production costs for a typical program and the documentation (i.e., the manual) that must accompany the program.

Development costs (fixed):			
Program development		$10,000	
Manual preparation and typesetting		3,000	
Advertising		$10,000	
Total			$23,000
Variable costs per unit:			
Blank disk	$2.00		
Loading cost	0.50		
Postage and handling	1.25		
Printing of the manual	$2.75		
Total			$6.50

A typical program of this type, including the manual, sells for $40. Based on this information:

1. Determine the breakeven number of programs and the total revenue associated with this volume.
2. MicroApplications has a minimum profit target of $40,000 on each new program it develops. Determine the unit and dollar volume of sales required to meet this goal.
3. While this program is still in the development stage, market prices for software fall by 25 percent due to a significant increase in the number of programs being supplied to the market. Determine the new breakeven unit and dollar volumes.

Solution

1. Based on fixed costs of $23,000, a price of $40 per unit, and variable costs per unit of $6.50, the unit volume required to break even is

$$Q_e = \frac{\text{fixed cost}}{P - AVC} = \frac{\$23{,}000}{\$40 - \$6.50} = 686.6$$

Total revenue at this output rate is determined by multiplying price times the breakeven quantity;

$$TR = PQ_e = 40(686.6) = \$27{,}464$$

2. The quantity necessary to meet the profit target of $40,000 is

$$Q_R = \frac{FC + \pi_R}{P - AVC} = \frac{23{,}000 + 40{,}000}{40 - 6.50} = 1{,}880.6$$

The associated total revenue is

$$TR = PQ_R = 40(1{,}880.6) = \$75{,}224$$

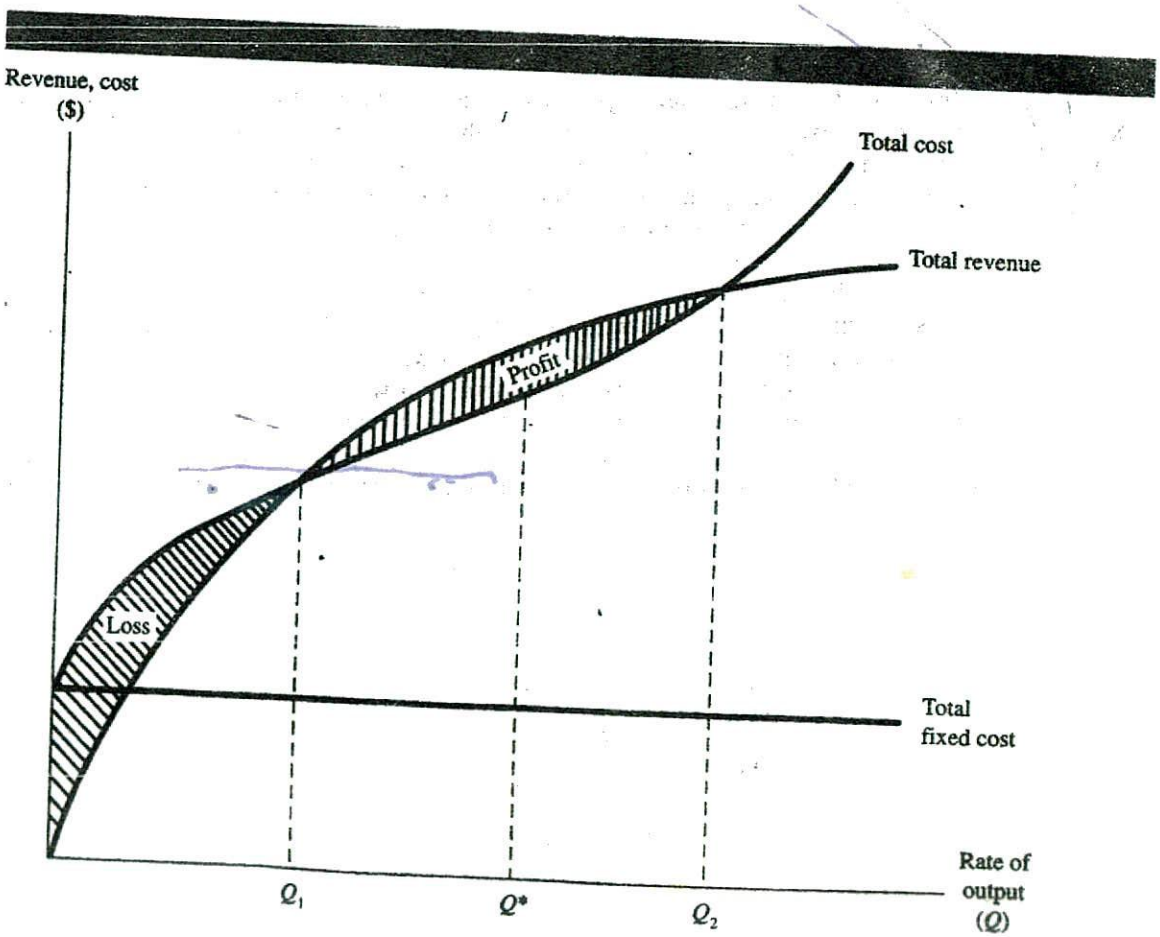
3. If the price declines by 25 percent to \$30, the new breakeven quantity would be

$$Q_e = \frac{23{,}000}{30 - 6.50} = 978.7$$

The corresponding total revenue for this output rate is

$$TR = 30(978.7) = \$29{,}361$$

If the assumptions of constant price and average variable cost are relaxed, breakeven analysis can still be applied, although the key relationships (total revenue and total variable cost) will not be linear functions of output. Nonlinear total revenue and cost functions are shown in Figure 7.7. The cost function is conventional in the sense that at first costs increase but less than in proportion to output and then increase more than in proportion to output. There are two breakeven points, Q_1 and Q_2. Note that profit, the vertical distance between the total revenue and total cost functions, is maximized at output rate Q^*.



Of the two breakeven points, only the first, corresponding to output rate Q_1, is relevant. When a firm begins production, management usually expects to incur losses. But it is important to know at what output rate the firm will go from a loss to a profit situation. In Figure 7.7, the firm would want to get to the breakeven output rate Q_1 as soon as possible and then, of course, move to the profit-maximizing rate Q^*. However, the firm would not expand production beyond Q^* because this would result in a reduction of profit. No rational manager would ever increase the rate of production to the second breakeven rate Q_2, and therefore, that point is irrelevant.

Key Concepts

- The output rate that yields a specified rate of economic profit is found by dividing the required profit plus total fixed cost by profit contribution, that is,

$$Q_R = \frac{\pi_R + TFC}{P - AVC}$$

- Breakeven analysis is a special case of required profit analysis where the required profit is zero.

Operating Leverage

A firm is said to be *highly leveraged* if fixed costs are large relative to variable costs. For example, relatively large fixed costs may result when the firm has large amounts of borrowed money with large fixed-interest obligations. Also, firms may have large investments in fixed assets, such as plant and equipment, with heavy fixed expenses, including depreciation charges and/or lease payments that must be paid regardless of the rate of output.

A general characteristic of a highly leveraged firm is that it experiences more variation in profits for a given percentage change in output than does a less leveraged firm. This is because total cost for a highly leveraged firm will change less than in proportion to a change in output rate, and therefore profit will tend to change more than in proportion to output changes. Conversely, for a firm with little or no fixed cost, profit should change more nearly in proportion to output changes.

Leverage can be analyzed using the concept of *profit elasticity* (E_π), defined as the percentage change in profit associated with a 1 percent change in unit sales or rate of output. That is,

$$E_\pi = \frac{\%\text{ change in profit}}{\%\text{ change in unit sales}} \tag{7-9}$$

or

$$E_\pi = \frac{\frac{\Delta\pi}{\pi}}{\frac{\Delta Q}{Q}} = \frac{\Delta\pi}{\Delta Q} \cdot \frac{Q}{\pi} \tag{7-10}$$

For infinitesimally small changes in Q, the profit elasticity is

$$E_\pi = \frac{d\pi}{dQ} \cdot \frac{Q}{\pi}$$

If the price of output is constant regardless of the rate of output, profit elasticity depends on three variables: the rate of output, the level of total fixed costs, and variable cost per unit of output. This can be seen by substituting the equations for profit

$$\pi = PQ - (AVC)(Q) - TFC$$

and change in profit

$$\Delta\pi = P(\Delta Q) - (AVC)(\Delta Q)$$

into equation (7-10). That is,

$$E_\pi = \frac{[P(\Delta Q) - (AVC)(\Delta Q)]/[PQ - (AVC)(Q) - TFC]}{\Delta Q/Q}$$

Simplifying this equation yields a computational formula for profit elasticity:

$$E_\pi = \frac{Q(P - AVC)}{Q(P - AVC) - TFC} \qquad (7\text{-}11)$$

Close inspection of this equation reveals that for two firms with equal prices, rates of output, and variable costs per unit, the firm having the greater total fixed cost will have the higher profit elasticity. This is because total fixed cost has a minus sign in the denominator in equation (7-11). The greater is total fixed cost, the smaller is the denominator and the higher is the value of E_π.

Table 7.5 shows profit and profit elasticity data for two firms, *A* and *B*, with differing total fixed costs and average variable costs. Note that leverage is greatest for smaller output rates and that it declines as output increases. At the same rate of output, the elasticity is always higher for firm *B* than for firm *A*. This is because *B* has higher total fixed costs and lower variable costs per unit than does *A*. Therefore, if both firms are producing the same output rate, for any change in output the percentage change in profit will be greater for *B* than for *A*. For example, at an output rate of 1,000, a 1 percent increase in output results in a 1.25 percent increase in profit for *A*, compared to a 2 percent increase for *B*. As the output rate increases, the difference in the elasticities between the two firms decreases.

TABLE 7.5 Operating Profit Elasticity for Two Firms

Firm *A*:
Price = $10.00
AVC = $ 5.00
TFC = $ 1,000

Firm *B*:
Price = $10.00
AVC = $ 2.00
TFC = $ 4,000

Rate of Output	Profit		Profit Elasticity	
	Firm A	Firm B	Firm A	Firm B
1,000	4,000	4,000	1.25	2.00
1,500	6,500	8,000	1.15	1.50
2,000	9,000	12,000	1.11	1.33
2,500	11,500	14,000	1.09	1.25
3,000	14,000	16,000	1.07	1.20

Over time, profits for firm *B* will vary considerably more than for *A*. In a sense, the management of *B* has structured the firm so that it takes more risk. When the rate of output is high, profits will be greater at *B* than at *A*. However, if economic conditions become unfavorable and output falls, profit will decline more rapidly for the highly leveraged firm *B*. If output continues to fall, firm *B* will incur losses before *A* will. In a prolonged period of low demand, the risk of bankruptcy is greater for *B*. Thus profit elasticity can be used as an indicator of risk.

As will be shown in chapter 14, there is usually a trade-off between risk and return. That is, greater returns are usually associated with greater risk, and vice versa. A management decision to become more leveraged (e.g., by incurring significant debt or a large capital investment program) is effectively a decision to accept greater risk for the chance to earn higher profit.

Key Concepts

- A firm is said to be highly leveraged if fixed costs are high relative to variable costs. In general, the use of leverage implies higher risk (i.e., more variability in profit over time).
- Leverage can be measured by profit elasticity, defined as the percentage change in profit associated with a 1 percent change in output.

ESTIMATING COST FUNCTIONS

As discussed in the previous chapter, many managerial decisions require quantitative information about the firm's production function. This is also true for the firm's cost functions. Decision making generally requires that the manager go beyond knowing that the average cost function is U-shaped and that marginal cost is increasing to actually making estimates of the parameters of these cost functions. In this section, methods for estimating both short-run and long-run cost functions are developed.

Short-Run Cost Functions

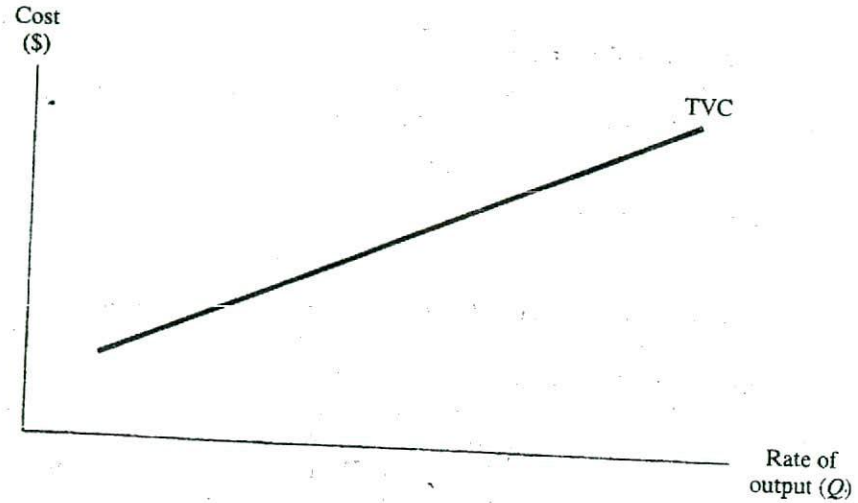
Recall that in the short run some costs are fixed. Although these fixed costs should be identified and measured, they typically are not used to estimate the short-run cost function. The usual procedure is to estimate the total or average variable cost function and then, if necessary, add the fixed cost component to obtain the total or average cost function.

Suppose that accurate data on variable cost and output have been collected. The remaining tasks are to specify the appropriate functional form (i.e., the hypothesized relationship between cost and output), statistically estimate the parameters of that functional form using the standard multiple-regression technique, and then interpret the results.

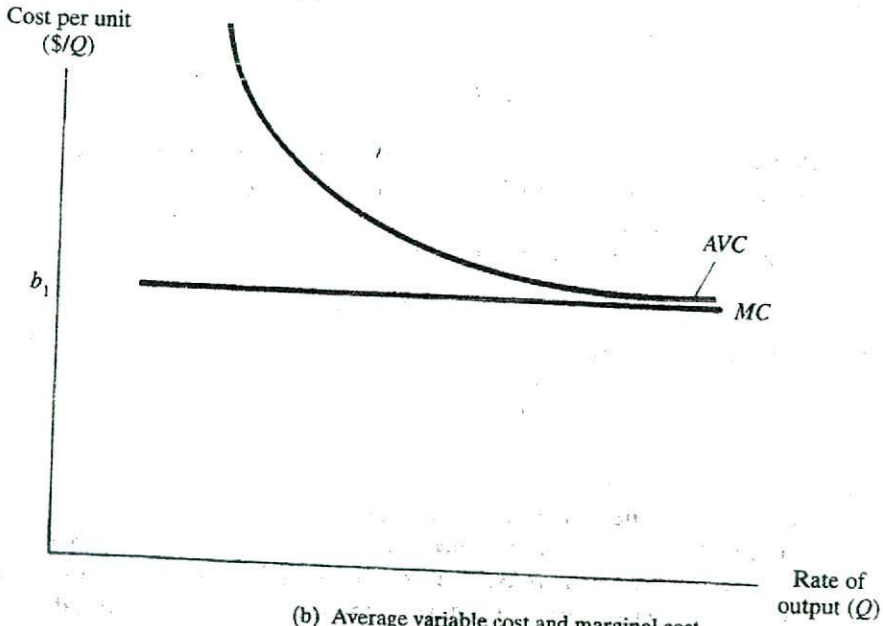
If the relationship between cost and output is approximately linear, the functional form

$$TVC = b_0 + b_1 Q \tag{7-12}$$

may be used to estimate the cost function. If b_0 and b_1 are estimated, the average variable cost function would be given by



(a) Total variable cost



(b) Average variable cost and marginal cost

FIGURE 7.8 Cost Curves Based on a Linear Total Cost Function

$$AVC = \frac{b_0}{Q} + b_1 \tag{7-13}$$

and the marginal cost function by

$$MC = b_1 \tag{7-14}$$

TVC, *AVC*, and *MC* curves that are consistent with this linear model are shown in Figure 7.8. These functions have the following properties: Total variable cost is a linear

function; average variable cost declines initially and then becomes quite flat approaching the value of marginal cost as output increases;[1] and marginal cost is constant at b_1.

If the empirical data indicate a U-shaped average cost curve, the linear function just used will not capture that relationship between output and cost. Consequently, a quadratic total variable cost function of the form

$$TVC = c_0 + c_1Q + c_2Q^2 \tag{7-15}$$

or a cubic cost function

$$TVC = d_0 + d_1Q + d_2Q^2 + d_3Q^3 \tag{7-16}$$

is often used because functions of this type can capture the hypothesized nonlinear relationship. The parameters of both (7-15) and (7-16) can be estimated directly using the least-squares regression method.[2]

Although the shape of the quadratic and cubic functions will depend on the estimated parameters, conventional or typical estimated cost functions based on these functional forms are shown in Figures 7.9 and 7.10. The quadratic function (Figure 7.9) has the following properties: Total cost increases at an increasing rate; marginal cost is a linearly increasing function of output[3] (i.e., $MC = c_1 + 2c_2Q$); and average variable cost, found by dividing the total variable cost function by Q, is a nonlinear increasing function, that is

$$AVC = \frac{c_0}{Q} + c_1 + c_2 = Q \tag{7-17}$$

Typical total variable cost, average cost, and marginal cost curves based on a cubic function are shown in Figure 7.10. The characteristics of these functions are: Total variable cost first increases at a decreasing rate (up to output rate Q_1 in the figure), then increases at an increasing rate; and both marginal cost

$$MC = d_1 + 2d_2Q + 3d_3Q^2 \tag{7-18}$$

and average variable cost

$$AVC = \frac{d_0}{Q} + d_1 + d_2Q + d_3Q^2 \tag{7-19}$$

are U-shaped functions.

In Figure 7.10, the conventional relationships among the cost functions are evident. For example, note the correspondence between the inflection point on TVC (i.e., at output Q_1, where the rate of increase goes from decreasing to increasing) and the minimum point on marginal cost. Also, marginal cost intersects the average variable cost function at the minimum point of the AVC curve.

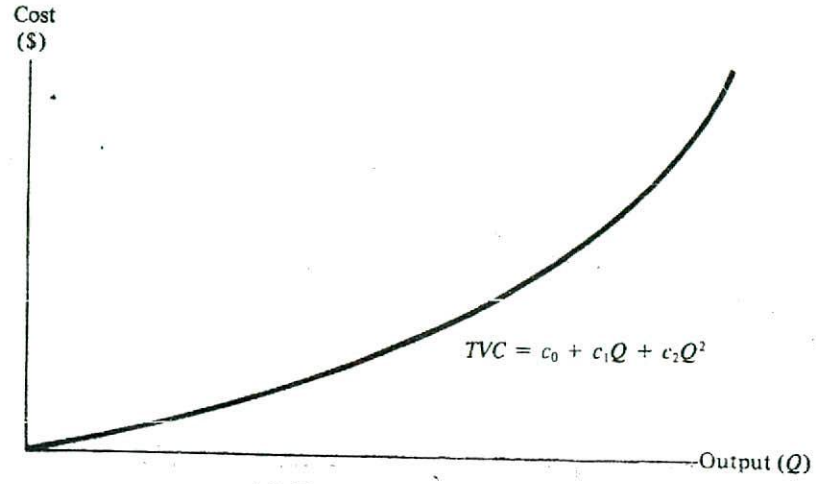
---

[1]Note that the ratio b_0/Q approaches zero as Q becomes large. Therefore, for high output rates, AVC will approach b_1, which is marginal cost.

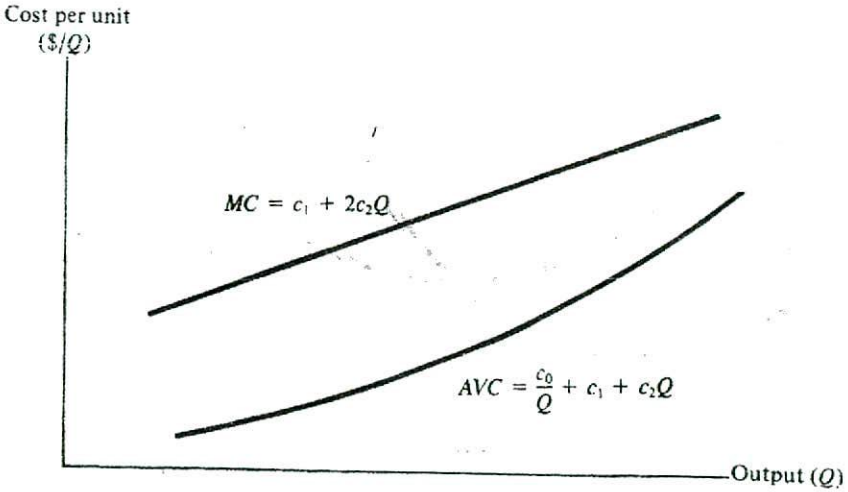
[2]Equation (7-16) is nonlinear in the variables Q, Q^2, and Q^3, but it is linear in the parameters d_0, d_1, d_2, and d_3. Therefore the ordinary last-squares regression method can be used without any transformation of the equation.

[3]Recall that marginal cost is the first derivative of the total variable cost function equation. That is,

$$MC = \frac{d(TVC)}{dQ} = c_1 + 2C_2Q$$



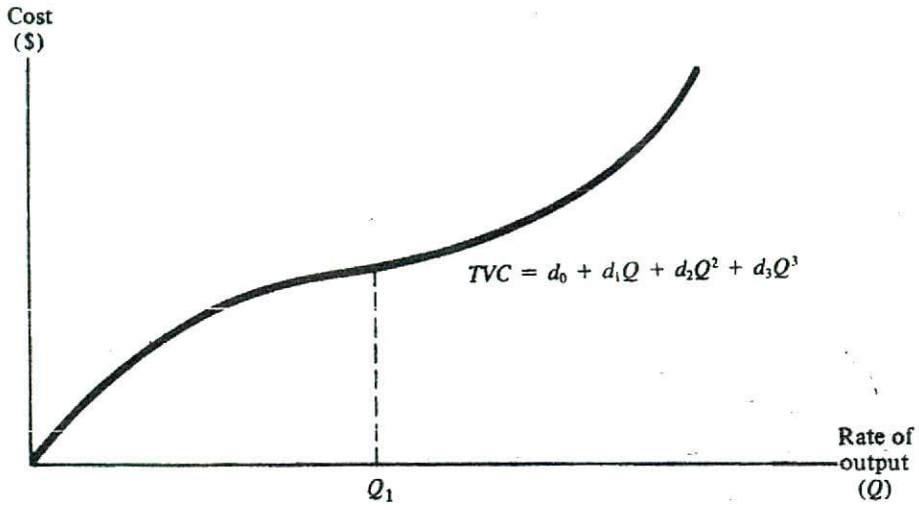
(a) Total variable cost



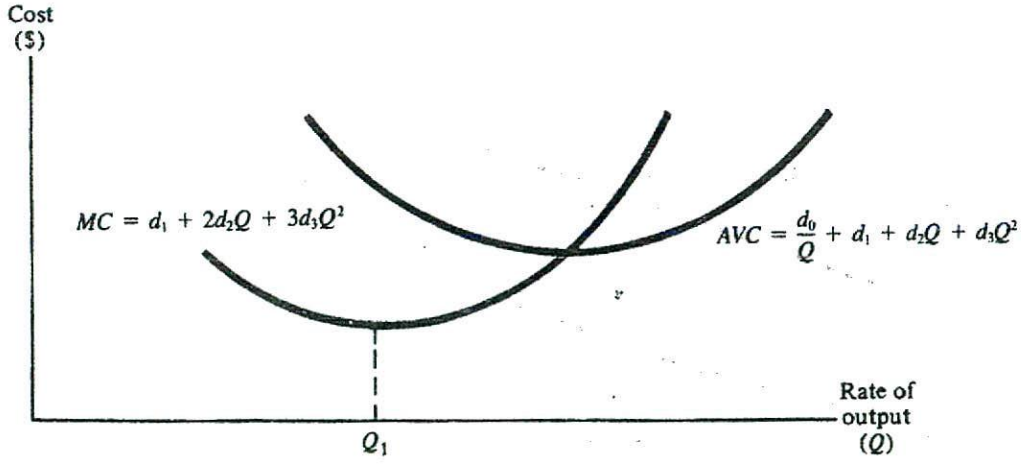
(b) Average variable cost and marginal cost

FIG 6 [illegible]

Although there is empirical evidence to support a variety of cost relationships, several studies of short-run cost functions have concluded that the marginal cost curve is approximately horizontal, that is, that marginal cost is constant over a fairly wide range of output rates. These empirical results would appear to be inconsistent with conventional cost theory, which suggests that average cost curves are U-shaped and that marginal cost functions are rising. This may be the result of saucer-shaped marginal cost functions that do have negative slopes at low rates of output and positive slopes at high output rates, but are essentially flat over fairly wide output ranges. As firms normally would be operating within that range (i.e., operating efficiently), all or most of the actual production rate and cost data would show approximately constant marginal costs.



(a) Total variable cost



(b) Average variable cost and marginal cost

An alternative explanation for a flat marginal cost function is that the so-called fixed inputs may not really be fixed. Recall from chapter 6 that the marginal product function declines because increasing amounts of the variable input are being combined with fixed amounts of another input. Initially, returns to additional units of the variable factor may increase, but ultimately the law of diminishing marginal returns applies, and the marginal product function declines while the marginal cost curve rises. But if the "fixed" input is not really fixed, the fixed-variable input proportion may not change as output varies, and the law of diminishing returns may not be observed. For example, a manufacturing firm may have a fixed stock of machines, but the number in use may vary

as the rate of output changes. In that case, the ratio of labor to machines in use may be roughly constant even though the ratio of labor to the total stock of machines may vary considerably. If this is the case, both the estimated marginal and average cost functions may be quite flat.

Example Estimating and Using Cost Functions

The engineering department of Consolidated Chemicals has developed the following output-cost data for a proposed new plant to produce ammonium sulfate fertilizer:

Output	*Total Cost*
50	870
100	920
150	990
200	1,240
250	1,440
300	1,940
350	2,330
400	3,100

a. Estimate the total cost function and then use that equation to determine the average and marginal cost functions. Assume a quadratic total cost function:

$$TC = c_0 + c_1Q + c_2Q^2$$

b. Determine the output rate that will minimize average cost and the per-unit cost at that rate of output.

c. The current market price of this fertilizer is \$5.50 per unit and is expected to remain at that level for the foreseeable future. Should the plant be built?

Solution

a. Using the ordinary least-squares regression method, the estimated function is

$$TC = 1{,}016 - 3.36Q + 0.021Q^2, \qquad R^2 = 0.99$$
$$(11.45)\ (-3.71)\ \ (10.71)$$

The *t*-statistics, shown in parentheses, indicate that the coefficients of each of the independent variables are significantly different from zero. The value for the coefficient of determination means that 99 percent of the variation in total cost is explained by changes in the rate of output.

The average cost function is

$$AC = \frac{TC}{Q} = \frac{1{,}016}{Q} - 3.36 + 0.021Q$$

and the marginal cost function is

$$MC = \frac{d(TC)}{dQ} = -3.36 + 0.042Q$$

b. The output rate that results in minimum per-unit cost is found by taking the first derivative of the average cost function, setting it equal to zero, and solving for Q.

$$\frac{d(AC)}{dQ} = -\frac{1{,}016}{Q^2} + 0.021 = 0$$

$$0.021 = \frac{1{,}016}{Q^2}$$

$$Q = 220$$

To find the cost at that rate of output, substitute 220 for Q in the average cost equation and solve.

$$AC = \frac{1{,}016}{220} - 3.36 + 0.021(220) - \$5.88$$

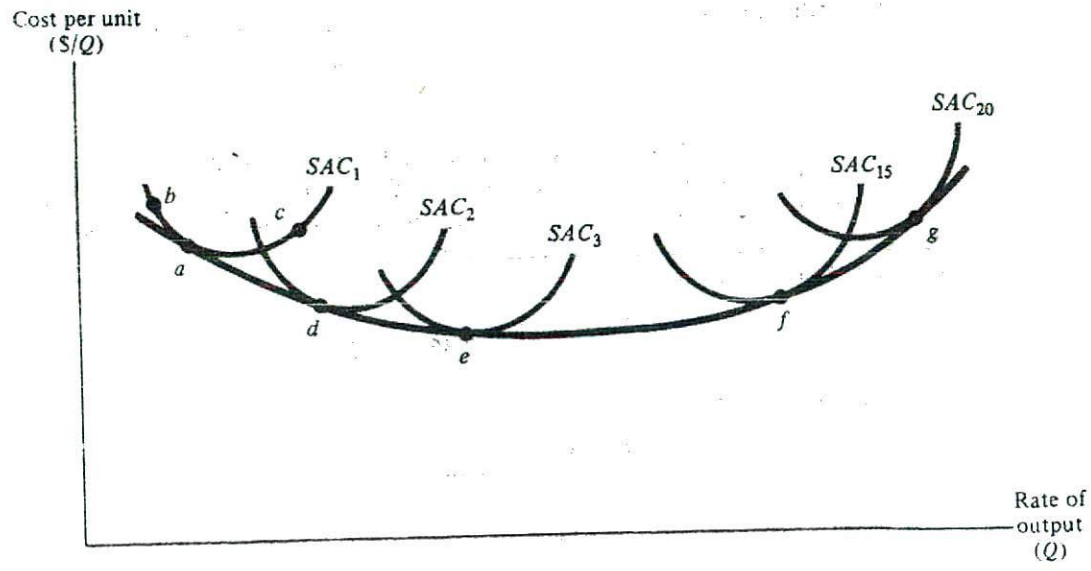
c. Because the lowest possible cost is \$5.88 per unit, which is \$0.38 above the market price, the plant should not be constructed.

Long-Run Cost Functions

Recall that the long-run average cost curve consists of points or small segments of each of a series of short-run average cost functions. A long-run cost function and associated short-run cost functions are shown in Figure 7.11. At any point in time, of course, a firm will be operating on one of these short-run functions. Estimating the long-run function requires: (1) obtaining data on each of a number of points on the long-run function (i.e., obtaining data on the relevant point on each of a number of short-run cost functions such as point *a, d, e, f,* and *g* on the short-run functions shown in Figure 7.11); (2) specifying the appropriate functional form; and (3) estimating the parameters of the function.

Either a time-series or a cross-section approach can be used, although most studies of long-run cost behavior have used cross-section data. One problem with time-series

FIGURE 7.11 Short-Run Average Cost Curves and the Associated Long-Run Average Cost Curve



data on a single plant is that if the period is long enough for the scale to have changed, it is probably long enough for production technology and input prices to have changed as well. As suggested previously, a cross-section study must assume that firms are operating at that point on the short-run function that lies on the long-run function. For example, in Figure 7.11 accurate estimation of the long-run cost function would require that the firm represented by SAC_1 would be operating at point *a*. If the firm were operating at some other point, such as *b* or *c*, the estimate of the long-run average cost curve would be biased in an upward direction.

Case Study

Returns to Scale in High School Education

Studies of per-student expenditures in the nation's schools have generally shown a tendency for these costs to increase as the number of students in the school increases. Such studies have led to the conclusion that there are decreasing returns to scale in the production of education. One problem with this approach is that it has not taken into account differences in the quality of output among schools. If the larger schools are producing a higher-quality product (i.e., a better-educated student), it is not legitimate to infer that higher cost per student in the larger schools is indicative of decreasing returns to scale. That is, the effect of quality differences must be held constant in order to estimate the relationship between average cost and school size.

Riew studied 102 accredited high schools in Wisconsin. He found that if adjustments are made for quality of education, there are significant economies of scale in producing high school education. Riew specified a general average cost function of the form:

$$\text{average cost} = f(\text{number of students, quality})$$

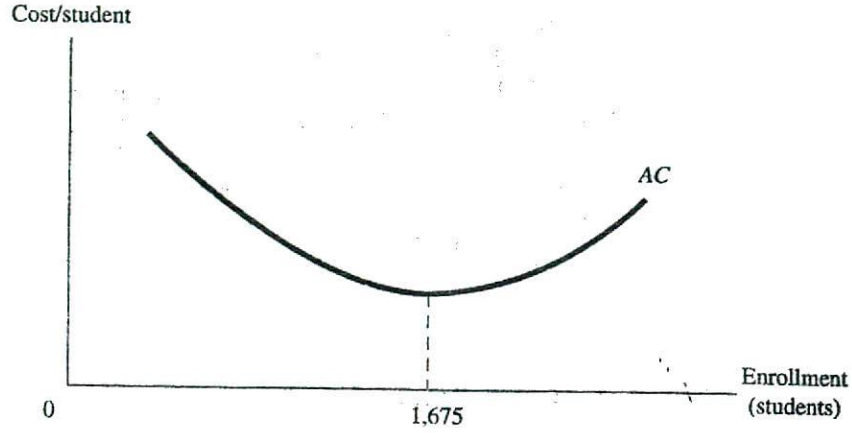
Average cost per student (AC) was measured by expenditure per student and number of students by average daily attendance. Although output quality generally is measured by reference to particular attributes of the good or service produced, in Riew's study quality was measured by the characteristics of the schools that were surveyed. The following measures of quality were used: percentage of classrooms built within the last 10 years; average teacher's salary; number of credit units offered; and average number of courses taught per teacher.

The least-squares regression technique was used to estimate the parameters of a quadratic cost equation of the form

$$AC = c_0 + c_1Q + c_2Q^2 + c_3V_1 + c_4V_2 + c_5V_3 + c_6V_4$$

where AC is cost per student, Q is number of students, and V_1, V_2, V_3, and V_4 are the quality variables. After adjusting for quality differences among schools so that the effect of variables V_1 through V_4 are included in the constant term, the following net relationship between cost and number of students was estimated:

$$AC = 10.3 - 0.402Q + 0.00012Q^2 \qquad R^2 = 0.56$$
$$(6.38) \quad (5.22)$$



Relationship between Cost per Student and Number of Students

The values of the *t*-statistics are shown in parentheses and indicate that both coefficients are significantly different from zero. The number of students that minimizes this function is found by setting the first derivative of the average cost function equal to zero and solving for Q. That is,

$$\frac{d(AC)}{dQ} = -0.402 + 0.00024Q = 0$$

$$Q = 1{,}675$$

Thus, average cost is minimized for a school with 1,675 students.* When graphed, this estimated function is a U-shaped average cost curve having a minimum point at 1,675 students, as shown in the above figure. ■

*The second derivative of the average cost function is positive (i.e., 0.00024); this implies that AC is minimized for a school of this size.

SOURCE: J. Riew, "Economies of Scale in High School Operations," Review of Economics and Statistics 48 (3): 280–287.

Key Concepts

- Managerial decisions often require that quantitative estimates be made of the firm's cost functions.
- A quadratic cost function of the form $TVC = c_0 + c_1Q + c_2Q^2$ yields positively sloped average and marginal cost curves.
- A cubic cost function of the form $TVC = d_0 + d_1Q + d_2Q^2 + d_3Q^3$ yields U-shaped average and marginal cost curves.

SUMMARY

The theory of cost is a fundamental concern of managerial economics. The best measure of resource cost is the value of that resource in its highest-valued alternative use (i.e., its opportunity cost). The concept of opportunity cost includes both explicit and implicit costs. Examples of the latter include the value of labor and capital contributed to the firm by the manager/owner.

The cost of a long-lived asset during the production period is the difference in the value of that asset between the beginning and end of the period. Because the opportunity cost concept is used, economic cost includes a normal return or profit to the firm. A normal return is defined as the minimum payment necessary to keep resources from moving to other firms or industries. Marginal and incremental costs are fundamental to decision making. Sunk costs generally are irrelevant.

The cost function relates cost to specific rates of output. The basis for the cost function is the production function and the prices of inputs. In the short run, the rate of one input is fixed. The cost associated with that input is called fixed cost. The cost associated with the variable inputs is defined as variable cost. The total cost of any rate of output is equal to the total fixed cost plus the total variable cost of producing that output rate. Per-unit cost functions include average total cost, average variable cost, average fixed cost, and marginal cost. Often, the per-unit cost functions are more useful for decision making than are the total cost functions.

In the long run, all costs are variable. The long-run average cost curve is the envelope of a series of short-run average cost curves. If returns to scale are increasing, long-run average cost (LAC) will be decreasing. If returns to scale are decreasing, LAC will be increasing. The long-run cost functions are used for planning the optimal scale of plant size.

Profit contribution analysis is used to determine the output rate necessary to earn a specified profit rate. A special case of profit contribution theory is breakeven analysis, where the rate of output necessary to generate a zero rate of economic profit is determined.

A firm is said to be highly leveraged if fixed costs are large relative to variable costs. Leverage is measured by the profit elasticity or the percentage change in profit associated with a 1 percent change in output. In general, increased leverage implies more variability in profit over time and therefore greater risk.

Linear, quadratic, and cubic functional forms are used to estimate short-run cost functions. Both the quadratic and cubic functions can be used as functional forms if the cost data are consistent with the nonlinear cost curves suggested by economic theory. Long-run cost functions typically are estimated using cross-section data on costs in a number of plants. It must be assumed that the cost-output data observed on the short-run function of each plant also is a point on the firm's long-run cost function.

Discussion Questions

7-1. Three business-school graduates decide to open a business, and all three devote their full time to its management. What cost would you assign to their time? Is this an explicit or implicit cost?

7-2. Why do increasing returns to scale imply a decreasing long-run average cost function and decreasing returns to scale imply an increasing long-run average cost function?

7-3. Why is the historic cost of inventory or capital equipment irrelevant for managerial decision making?

7-4. Virtually all firms report the value of capital equipment and real estate based on historical cost less accounting depreciation. Do you think the reported value of these assets would approximate their market value? Explain. Would reporting the true market value of these assets provide investors with better information? Why?

7-5. If a firm found it could only operate at a breakeven output rate, would it stay in business in the long run?

7-6. Is there a problem using accounting data based on historical cost and standard depreciation techniques to estimate economic cost functions? Explain.

7-7. Some empirical studies have suggested that the marginal cost function is approximately horizontal, but conventional cost theory suggests that the marginal cost curve is U-shaped. Provide an explanation for this apparent inconsistency.

7-8. A pharmaceutical firm has spent $5 million developing and testing a new antibaldness drug. The head of the marketing department now estimates that it will cost $3 million in advertising to launch this new product. Total revenue from all future sales is estimated at $6 million, and therefore, total costs will exceed revenue by $2 million. He recommends that this product be dropped from the firm's product offerings. What is your reaction to this recommendation? The head of the accounting department now indicates that $3.5 million of corporate overhead expenses also will be assigned to this product if it is marketed. Does this new information affect your decision? Explain.

Problems

7-1. Use the following data to compute total variable cost, total fixed cost, average total cost, average variable cost, average fixed cost, and marginal cost for each output rate shown. Also, use the information to determine equations for each of the total and per-unit cost functions.

Production Period	*Rate of Output (Q)*	*Total Cost (TC)*
1	10	1,800
2	0	1,000
3	4	1,320
4	2	1,160
5	7	1,560

7-2. Wyatt Motors has one fixed input, the long-term lease on its factory building, for which the rent is $5,000 per production period. Use the data shown here to determine average cost, average variable cost, and marginal cost for each output rate shown. Also, write equations for total cost, total variable cost, and marginal cost.

Q	*Total Variable Cost*
1	$1,000
2	2,000
3	3,000
4	4,000
5	5,000

7-3. Based on a consulting economist's report, the total and marginal cost functions for Advanced Electronics, Inc. are

$$TC = 200 + 5Q - 0.04Q^2 + 0.001Q^3$$
$$MC = 5 - 0.08Q + 0.003Q^2$$

The president of the company determines that knowing only these equations is inadequate for decision making. You have been directed to do the following:

a. Determine the level of fixed cost (if any) and equations for average total cost, average variable cost, and average fixed cost.
b. Determine the rate of output that results in minimum average variable cost.
c. If fixed costs increase to $500, what output rate will result in minimum average variable cost?

7-4. Given the following total revenue and total cost functions

$$TR = 50Q$$
$$TC = 10{,}000 + 30Q$$

a. Determine the breakeven rate of output.
b. Determine the output rate necessary to earn a profit of $20,000.

7-5. A firm has total fixed costs of $20 and sells output at $10 per unit. Profit contribution is 20 percent of price.

a. What is the breakeven rate of output?
b. What is the profit elasticity at an output rate of 20?

7-6. President Emert of Eastern State University had decided on a new budgeting system for all departments in the university. Historically, each department was provided with an annual budget that was essentially the same amount each year. Being a trained economist, the president has decided to allow student demand for courses to determine how the University's budget is to be allocated to departments. Each semester, every department will receive $40 for each student credit hour (SCH) taught. For example, a three-credit course for fall semester might enroll thirty-five students, thus generating 105 SCH. For offering that course, the department would receive $4,200 (i.e., $40 × 105 SCH).

The following table shows selected cost data for each of four university departments:

Department	*Number of Faculty*	*Salary Cost (Fixed)*	*Variable Cost per SCH*
Economics	10	$500,000	$3.00
English	15	510,000	2.00
Physics	8	380,000	7.00
Physical education	7	200,000	3.00

a. Determine the breakeven number of student credit hours for each department.
b. If the objective of each department is to maximize the size of its budget, what change in the type of course offerings and assignment of faculty to courses might result from such a budgeting system?
c. Suppose that student credit hours have been declining steadily for several years in physical education, while they have been growing rapidly in economics.

Given these trends, what advantages and disadvantages would the new budgeting system offer compared to the one previously used?

7-7. Three firms in the same industry all sell their product at $20 per unit. Their total fixed cost and average cost per unit are shown below:

	A	*B*	*C*
Total Fixed Cost	$200	$500	$1,000
Average variable cost	15	10	5

a. What is the breakeven output rate for each firm?
b. What is the profit elasticity for each firm at an output rate of 200?
c. Which firm is the most leveraged? Which is the least leveraged?

7-8. Interglobal Life Insurance arbitrarily assigns $400,000 per year of corporate fixed overhead expense to each local office. The manager of the Southern Region is considering opening an office in Baton Rouge, Louisiana. She estimates that revenue for this office would be $2,000,000 and expenses would be:

Rent	50,000
Labor	600,000
Taxes	50,000
Utilities	60,000
Advertising	250,000

Should the office be opened? (That is, can a profit be earned?)

7-9. Ruby Vazquez has invested $80,000 in a hardware store. Business has been good, and the store shows an accounting profit of $10,000 for the last year. This profit is after taxes and after payment of a $20,000 salary to Ms. Vazquez. This salary is less than the $40,000 she could make at another job. Considering the risk involved in the hardware business, she believes that a 15 percent after-tax rate of return is appropriate for this type of investment.

a. Given this information, calculate the *economic* profit earned by Ms. Vazquez.
b. What accounting profit would the firm have to earn in order for the firm to break even in terms of economic profit?

7-10. Universal Dental Products, Inc. manufactures false teeth with a pliable base that allows one size to fit any mouth. A set of these dentures sells for $80. Fixed costs are $200,000 per production period and the profit contribution is 40 percent of price.

a. Determine the profit elasticity at output rates of 8,000, 10,000, and 12,000 units.
b. For the next production period, fixed costs will increase to $300,000 due to a major capital investment program, but the new and more efficient machinery will result in lower variable production costs so that variable cost per unit will be reduced by $8. If price is unchanged, recompute the profit elasticity at output rates of 8,000, 10,000, and 12,000 units.
c. What change in the risk–return trade-off has the company made by this capital investment program?

7-11. Space Dynamics, Inc. produces electronic components for an antimissile system. Each component sells for $900, average variable cost is constant at $700, and total fixed cost is $10,000.
   a. Determine the breakeven rate of output.
   b. Demonstrate that the profit elasticity declines as the output rate increases.
   c. Show that for any rate of output, profit elasticity increases if fixed cost increases and this elasticity will decrease as variable cost per unit decreases.

7-12. Southern Airways, a small regional airline, has a daily late evening flight into Atlanta. The plane must be in Atlanta at 8:00 A.M. each morning for a flight to Richmond, Virginia. Unfortunately, the charge for a plane remaining overnight in Atlanta is $500. One alternative is to schedule a 9:00 P.M. flight to Gainesville, Florida (a one-hour trip), and a 6:00 A.M. flight back to Atlanta. There is no charge for the overnight stay at Gainesville. However, the flights to and from Gainesville will average only about ten passengers each at a one-way fare of $50. The operating cost of the plane is $600 per hour. The company's total fixed costs (which are unaffected by the decision to be made here) are allocated to each flight at the rate of $300 per hour. Should the plane stay in Atlanta overnight, or should flights to and from Gainesville be scheduled? Explain your decision.

7-13. A firm is considering the rental of a new copying machine. The rental terms of each of the three machines under consideration are given here:

	Costs		
Machine	*Monthly Fee*		*Per Copy*
A	$1,000	+	$0.03
B	300	+	0.04
C	100	+	0.05

How many copies per month would the firm have to make for *B* to be a lower total cost machine than *C*? For *A* to be lower cost than *B*?

7-14. During the last period, the sum of average profit and fixed costs for a firm totaled $100,000. Unit sales were 10,000. If variable cost per unit was $4, what was the selling price of a unit of output? How much would profit change if the firm produced and sold 11,000 units of output? (Assume average variable cost remains at $4 per unit.)

7-15. The economics department of Western Drilling, a producer of natural gas, has estimated the long-run total cost function for natural gas distribution to be

$$TC = 200Q - 0.004Q^3$$

where TC is total cost and Q is millions of cubic feet (MMCF) of natural gas per day.
   a. Determine an equation for the long-run average cost of distributing natural gas and plot it on a graph over the range $10 \leq Q \leq 150$.
   b. At present, Western produces but does not distribute natural gas. Interstate Pipeline is the only distributor of gas in the region, and it carries about 100 MMCF per day. Management at Western estimates the regional market will grow from 100 to 150 MMCF per day and thinks it might be able to capture about 50 percent of the increase in the size of the market. Interstate has the

capacity to deliver 200 MMCF per day. Will Western be able to compete against Interstate in the distribution of gas? That is, will Western be able to earn a normal return at the output rate associated with that part of the market it expects to capture? Explain. (Assume Interstate has the same total cost function as Western Drilling.)

Problems Requiring Calculus

7-16. Economists at Jensen Enterprises used time-series data to estimate the following total cost function for the firm:

$$TC = 200 - 2Q + 0.05Q^2$$

where TC is total cost and Q is the output rate.

a. Determine an equation for the average cost function. Plot this function, and find the output rate that minimizes average cost.
b. Is the production process characterized by decreasing, constant, or increasing returns to scale?
c. If the market price of output is \$4.32 per unit, is there a scale of plant that would allow the firm to earn an economic profit or at least to break even?

7-17. Given the total cost function for Randle Enterprises:

$$TC = 100Q - 3Q^2 + 0.1Q^3$$

a. Determine the average cost function and the rate of output that will minimize average cost.
b. Determine the marginal cost function and rate of output that will minimize marginal cost.
c. At what rate of output does average cost equal marginal cost?

7-18. Logan Manufacturing produces ballpoint pens. Fixed costs in each production period are \$25,000, and the total variable cost (TVC) is given by the equation

$$TVC = 0.15Q + 0.1Q^2$$

where Q is the rate of output. What output rate would minimize average total cost?

7-19. Given the total cost function

$$TC = 1{,}000 + 200Q - 9Q^2 + 0.25Q^3$$

a. Determine the equation for each total and per-unit cost function (i.e., TVC, FC, AFC, AVC, AC, and MC).
b. Determine the lowest price for output that would allow the firm to break even.
c. Determine the lowest price for output that would allow the firm to cover average variable cost.

7-20. A firm sells its output for \$20 per unit and has a total cost function

$$TC = 16 + 17Q - 9Q^2 + Q^3$$

a. Determine the firm's total profit function.
b. Determine the firm's marginal cost function.
c. Determine the profit elasticity at an output rate of 8 units.

Computer Problems

The following problems can be solved by using the TOOLS program (downloadable from www.prenhall.com/petersen) or by using other computer software.

7-21. The quantity produced (in thousands) and the average cost (in dollars) of producing a toy at different plants are shown here.

Plant	*Average Cost*	*Quantity*	*(Quantity)²*
1	$0.75	100	10,000
2	0.40	200	40,000
3	0.50	140	19,600
4	0.60	260	67,600
5	0.45	160	25,600
6	0.55	120	14,400
7	0.70	280	78,400
8	0.45	180	32,400
9	0.40	220	48,400
10	0.45	240	57,600

a. Use regression analysis to estimate average cost as a linear function of quantity produced. Write the equation, *t*-statistics, and coefficient of determination. Does the equation exhibit increasing, decreasing, or constant returns to scale?

b. Use regression analysis to estimate average cost as a linear function of quantity and quantity squared (e.g., the quantity and quantity squared data for plant 1 would be 100 and 10,000, respectively). Determine the equation, *t*-statistics, and coefficient of determination. At what quantity is average cost a minimum? Over what range of output are increasing returns to scale indicated? What about decreasing returns to scale?

c. Using the results from part (b), what is the minimum output necessary to break even at a price of $0.55?

7-22. A plant has been operating for 10 periods. The output rate and total cost for each period are shown here.

Period	*Q*	*TC*
1	4	$ 1,300
2	12	7,700
3	21	4,400
4	16	2,900
5	30	10,500
6	11	1,900
7	6	1,250
8	27	9,500
9	19	3,550
10	8	1,400

a. Use the multiple regression technique to estimate the cubic total cost function:

$$TC = d_0 + d_1Q + d_2Q^2 + d_3Q^3$$

(Hint: Create two additional variables, Q^2 and Q^3, and then regress TC on Q, Q^2, and Q^3.)

b. Use the estimated total cost function to derive equations for average cost and marginal cost.

c. Using the same data, compute average costs by dividing TC by Q for each of the 10 periods. Using the computed average cost data, estimate the quadratic average cost function.

$$AC = e_0 + e_1Q + e_2Q^2$$

How do your results compare with the estimated average cost equation from part (b)?

CHAPTER

8 Linear Programming

- **Preview**
- **Linear Programming Applications**
- **The Linearity Assumption**
- **Constrained Profit Maximization**
  - Structuring the Problem
  - The Feasible Region
  - Graphic Solution
  - Algebraic Solution
- **Constrained Cost Minimization**
  - Structuring the Problem
  - Algebraic Solution
  - Sensitivity Analysis
- **Special Problems in Linear Programming**
  - Multiple Solutions
  - Redundant Constraints
  - No Feasible Solution
- **The Dual Problem**
  - Structuring the Dual Problem
  - Solving the Dual Problem
- **Summary**
- **Discussion Questions**
- **Problems**

PREVIEW

In chapter 6, the concept of a constrained optimization problem was introduced. In general, such problems consist of an objective function and one or more constraints. For example, one problem addressed in that chapter is to minimize cost (the objective) subject to meeting a specified level of production (the constraint). The essence of economics is efficient resource allocation when resources are scarce. Specifying one or more constraints is simply a way of identifying this scarcity.

Linear programming (LP) is a technique for solving a special set of constrained optimization problems where the objective function is linear and there are one or more linear constraints. It is a powerful decision-making technique that has found application in a variety of managerial problems. Not only does linear programming have value in decision making, it also helps in understanding the concepts of constrained optimization and opportunity cost.

In the first section of this chapter, examples of problems that lend themselves to linear programming analysis are suggested. In the next two sections, different types of programming problems are considered. First, a profit-maximization problem is set up and solved and then a cost-minimization problem is considered. In each case, both a graphic and an algebraic approach to the solution are presented. Actually, most real-world programming problems are solved by computers, but it is important to understand the principles underlying the solutions. The section on sensitivity analysis shows how changes in one or more parameters of the linear programming problem affect the final solution. In the fifth section, a set of special considerations relevant to linear programming problems is presented. Finally, the concept of primal and dual linear programs is discussed. There it is shown how a value can be assigned to each of the resources constraining the objective function.

LINEAR PROGRAMMING APPLICATIONS

Because all managers face constrained optimization problems, it should not be surprising that linear programming has many uses in business. The earliest applications were in production-related problems. For example, managers in multiproduct firms sought the combination of output rates for each of several products that would maximize profit subject to a limited number of machine-hours and worker-days. An alternative production problem might have been to find the combination of output rates that could be produced at minimum cost while meeting a specified set of orders and inventory requirements.

In the marketing area, linear programming is used to select the minimum-cost mix of radio, television, and magazine advertising that will meet constraints on the total number of people exposed to the advertising and the number in certain age, sex, and income classes. Financial managers use linear programming models to determine the least-cost method of financing the firm given such constraints as bank borrowing limitations and a maximum ratio of debt to equity. In addition, the technique is now being used to value leases, bonds, pension liabilities, and options on common stock.

Many transportation problems can also be solved by linear programming. Consider an automobile producer with several manufacturing plants and numerous dealers throughout the country. How should shipments of cars be made from plants to dealers

to minimize total transportation costs? A special transportation variant of linear programming has been developed to solve problems of this type. Also, firms in the airline industry use the technique to efficiently assign crews to flights.

Linear programming has been widely used in agriculture. For example, dairy farmers have used the technique to determine the least-cost combination of feeds that would meet minimum nutrition requirements for their animals. Agribusiness managers also have used linear programming to determine the profit-maximizing allocation of land to different crops subject to such constraints as fixed land area, differing soil characteristics, and limited water availability.

These are but a few of the many constrained optimization problems that face managers. Historically, some managers used intuition and guesswork to make resource allocation decisions. In the era of scientific management, however, better techniques are available and should be used if the firm is to compete successfully. Linear programming is among the most important of these techniques.

THE LINEARITY ASSUMPTION

All the important relationships in a linear programming problem must be linear. These often include the production, total cost, total revenue, and profit functions. The assumption that these functions are linear also means it is assumed that there are constant returns to scale in production, that the price of output is constant for all output levels, and that input prices are constant regardless of the amount of input purchased. The combination of constant returns to scale and constant input prices implies that production cost per unit of output is constant. Further, the combination of constant per-unit cost and output price results in a constant rate of profit per unit of output.

Some critics of linear programming argue that the assumption of linearity is inconsistent with other aspects of economic theory that are based on U-shaped cost curves, production functions having other than constant returns to scale, and nonlinear profit functions. Although these critics have a point, cost and profit per unit may be constant over a limited range of output, and thus the assumptions underlying linear programming would be valid for that range. More important, the successful application of linear programming to many managerial problems suggests that the technique is a useful management tool. Even where the important relationships are not exactly linear, the programming approach still may be the best way to approximate optimal resource allocation.[1]

Key Concepts

- Linear programming is a widely used quantitative technique for solving constrained optimization problems where both the objective function and the constraints are linear.
- The linearity assumption means that there are constant output and input prices, constant returns to scale in production, and constant cost and profit per unit of output.

[1]There are nonlinear programming methods for problems involving nonlinear functions. For example, see D. Luenberger, *Linear and Nonlinear Programming* (Reading, MA: Addison-Wesley, 1989).

CONSTRAINED PROFIT MAXIMIZATION

Many firms simultaneously produce several products in a plant. Management faces the problem of whether to produce more or less of one product (i.e., to allocate more or less resources to that product) relative to the others. That decision is based on comparing the profit earned on an additional unit of output to the opportunity cost of the resources devoted to producing that unit. Linear programming is well suited to solving problems of this type.

Structuring the Problem

Consider a firm that produces two products, A and B, that require processing on three different machines. During the production period, the number of hours available on each machine is limited. Assume that profit per unit of each output is constant for all relevant rates of output. The problem facing the firm is to determine the quantities of A and B (i.e., Q_A and Q_B) that will maximize profit but not require more than the limited number of machine hours available. Q_A and Q_B are the *decision variables* in this problem.

Management's objective must be stated in the form of a function, hence, the term *objective function.* In this problem the objective function is

$$\pi = aQ_A + bQ_B \tag{8-1}$$

where π represents total profit and a and b are the profit per unit of A and B produced. Solving for Q_B yields

$$Q_B = \frac{\pi}{b} - \frac{a}{b}Q_A \tag{8-2}$$

Equation (8-2) describes a straight line where the vertical intercept is determined by the ratio of total profit to profit per unit of product B (i.e., π/b) and the slope is the negative of the ratio a/b, or the relative profitability of the two products. As both a and b are positive, the slope of the line will be negative.

Suppose that profit per unit of output is \$3 for product A and \$1 for product B. The profit or objective function is then

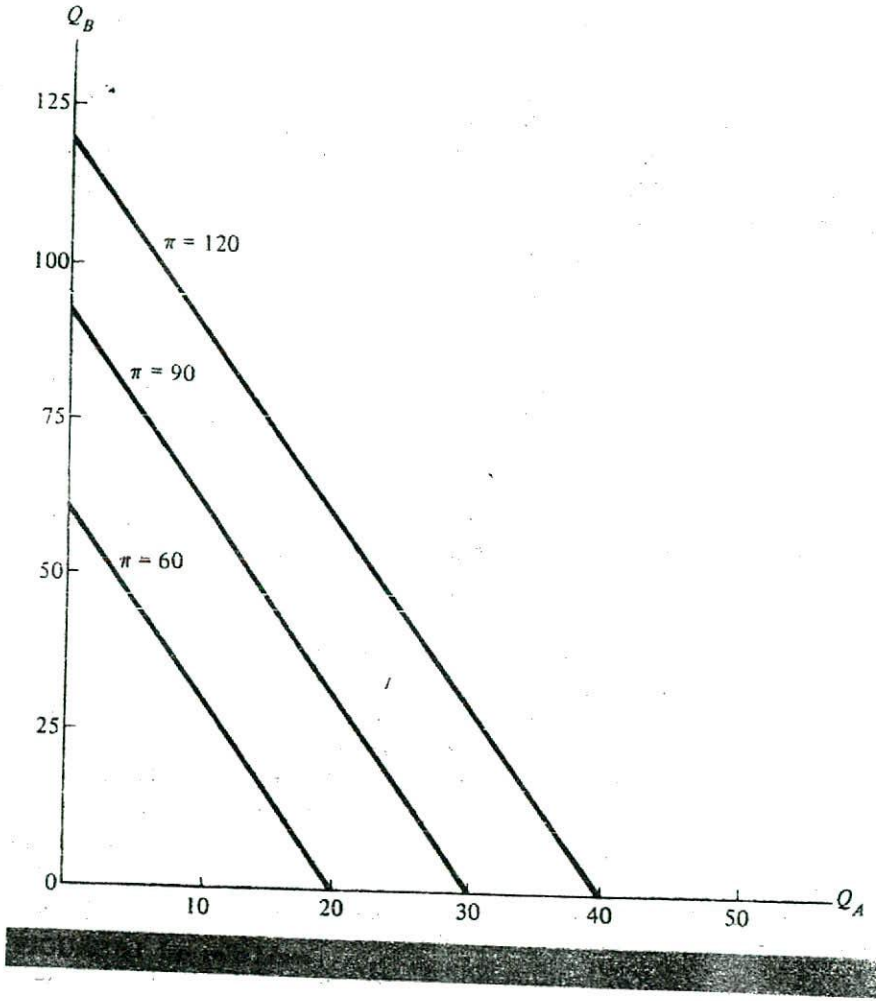
$$\pi = 3Q_A + 1Q_B$$

which can be rewritten as

$$Q_B = \pi - 3Q_A \tag{8-3}$$

If the rate of profit, π, is specified, equation (8-3) defines all combinations of Q_A and Q_B that yield that rate of profit. Thus equation (8-3) can be considered an isoprofit equation. For example, if $\pi = 90$, some of the combinations of Q_A and Q_B that yield that profit rate are

Q_A	Q_B
0	90
10	60
20	30
30	0



This \$90 isoprofit line is shown in Figure 8.1 together with the isoprofit lines for $\pi = 60$ and 120. The management problem is to attain the highest profit possible given the resource constraints facing the firm.

Recall that producing products A and B requires processing on three different machines. Let these machines be designated X, Y, and Z. Table 8.1 lists the hours of time

Machine	Hours of Machine Time Required Per Unit of Output for		Total Hours of Machine Time Available
	Product A	Product B	
X	1	3	90
Y	2	2	80
Z	2	0	60

per unit of output that are required on each machine and the hours of time available on each machine during the production period.

For example, on machine X each unit of product A requires one hour and each unit of B requires three hours. Thus, for any output rate, Q_A, the number of machine X hours required is $1Q_A$; for any output rate, Q_B, the number of hours required on machine X is $3Q_B$. Thus the total hours required on machine X is $1Q_A + 3Q_B$, and Table 8.1 indicates that this total is limited to no more than ninety hours during the production period. Therefore, the constraint for machine X is written

$$1Q_A + 3Q_B \leq 90 \tag{8-4}$$

This inequality reflects the limited resource availability of machine time.

Both products require two hours of time per unit of output on machine Y and the total time available is 80 hours. Thus the constraint for machine Y is

$$2Q_A + 2Q_B \leq 80 \tag{8-5}$$

Finally, product A requires two hours of time on machine Z per unit of output, but product B does not require the use of this machine, so the constraint is

$$2Q_A \leq 60 \tag{8-6}$$

In addition, a set of nonnegativity requirements is added to assure that the solution makes economic sense. These requirements specify that all decision variables (e.g., output of each product) be positive. Negative values for output would be meaningless, but without these nonnegativity constraints, the mathematical approach could easily result in one or more values of the decision variables being negative. For example, because profit per unit of output is higher for A than for B, if the nonnegativity requirements were not imposed, a solution might result in a large positive value for Q_A and a large negative value for Q_B.

Requiring that both Q_A and Q_B be nonnegative (i.e., $Q_A \geq 0$ and $Q_B \geq 0$) also ensures that the hours of machine time will also be nonnegative. The hours of time on each machine are specified by constraints (8-4), (8-5), and (8-6). The only way that negative machine hours could occur would be for there to be negative values for one or both of the output rates Q_A and Q_B.

Now the linear programming problem is complete and can be stated as follows:

maximize:	$\pi = 3Q_A + 1Q_B$	(objective function)
subject to:	$1Q_A + 3Q_B \leq 90$	machine X constraint
	$2Q_A + 2Q_B \leq 80$	machine Y constraint
	$2Q_A \leq 60$	machine Z constraint
	$Q_A \geq 0$	nonnegativity constraints
	$Q_B \geq 0$	

In general, the number of nonzero decision variables will be no greater than the number of resource constraints (i.e., the number of constraints excluding the nonnegativity constraints).

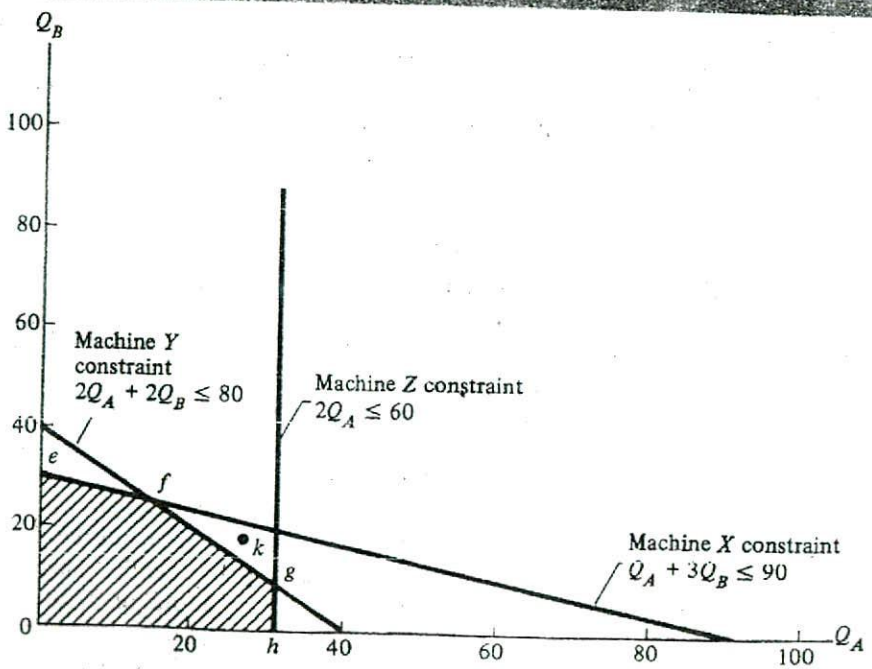
The Feasible Region

The next step is to determine the set of output rates Q_A and Q_B that can be produced without violating the constraints. The five constraints are shown graphically in Figure 8.2. Consider the first constraint, $Q_A + 3Q_B \leq 90$, which is shown as the machine X constraint in the figure. To draw this constraint, consider the relation as an equality and solve for Q_B, that is,

$$Q_B = 30 - \frac{1}{3}Q_A$$

This is the equation for a straight line that has a vertical intercept of 30 and a slope of $-1/3$. Now find two points on that line by setting $Q_A = 0$ and solving for $Q_B = 30$, yielding the point (0, 30), and then by setting $Q_B = 0$ and solving for $Q_A = 90$, thus yielding a second point (90, 0). These two points are sufficient to identify the line. Because it is a less-than-or-equal-to constraint, all values on or below this line meet the constraint. Similarly, all points on or below the line $2Q_A + 2Q_B \leq 80$ meet the machine Y constraint, and points to the left of $2Q_A \leq 60$ meet the constraint on hours available for machine Z. Finally, the nonnegativity requirements $Q_A, Q_B \geq 0$ restrict the allowable rates of the decision variables to the northeast quadrant, that is, where Q_A and Q_B are nonnegative.

In general, the feasible region consists of all values of the decision variables that satisfy all the constraints simultaneously. In this example, the feasible region of production is that set of combinations of the decision variables, Q_A and Q_B, that meets all



five constraints. This set is shown as the shaded area in Figure 8.2. Any point on or within that boundary can be produced because all the constraints are satisfied. A point such as k is not in the feasible region because it does not satisfy the constraint for machine Y. The linear programming problem is to identify the one combination (Q_A, Q_B) within the feasible region that yields maximum profit.

Key Concepts

- All linear programming problems consist of an objective function, one or more resource constraints, and a nonnegativity constraint for each decision variable.
- The feasible region consists of all combinations of values for the decision variables that satisfy all of the constraints.
- The goal of linear programming is to find that point in the feasible region that optimizes (i.e., either maximizes or minimizes) the objective function.

Graphic Solution

Recall the problem of maximizing production subject to a budget constraint from chapter 6. The budget constraint defined a feasible set of input combinations. The problem was to find the highest isoquant that touched that budget constraint. The linear programming problem discussed here is analogous to that production problem. Here the problem is to find the highest isoprofit line that still meets the constraints, that is, to find the highest isoprofit line that has at least one point in common with the feasible region.

In Figure 8.3, the feasible region is shown together with three isoprofit functions, which are depicted as dashed lines. The highest profit line that satisfies all of the constraints is $\pi = 100$, which touches the feasible region at point g. The coordinates of point g ($Q_A = 30$, $Q_B = 10$) provide the solution to this profit-maximizing problem. Substituting these values into the profit function confirms that profit is \$100 at this point:

$$\pi = 3(30) + (10) = 100$$

Any higher profit line, for example, that for \$150 of profit, would have no point in common with the feasible region. There are feasible combinations of Q_A and Q_B on profit lines below the \$100 isoprofit line, such as point j on the \$60 isoprofit line, but they are suboptimal because they imply lower profits.

The machine time allocated to each product is determined by multiplying the solution values Q_A and Q_B by the machine time requirements specified in Table 8.2. For example, for machine X, the number of hours used is 30 + 3(10), or 60 hours. The number of hours of time on each machine allocated to each product is shown in Table 8.2. Note that the available hours for machines Y and Z are all used, but there are unused hours available for machine X. Producing 30 units of A requires 30 hours on machine X, and producing 10 units of B also requires 30 hours, for a total of 60 hours required. As there are 90 hours available, 30 hours remain unused. In the language of linear programming, it is said that the constraints for machines Y and Z are binding, but there is slack in time for machine X. The constraint for X is said to be nonbinding. Note the op-

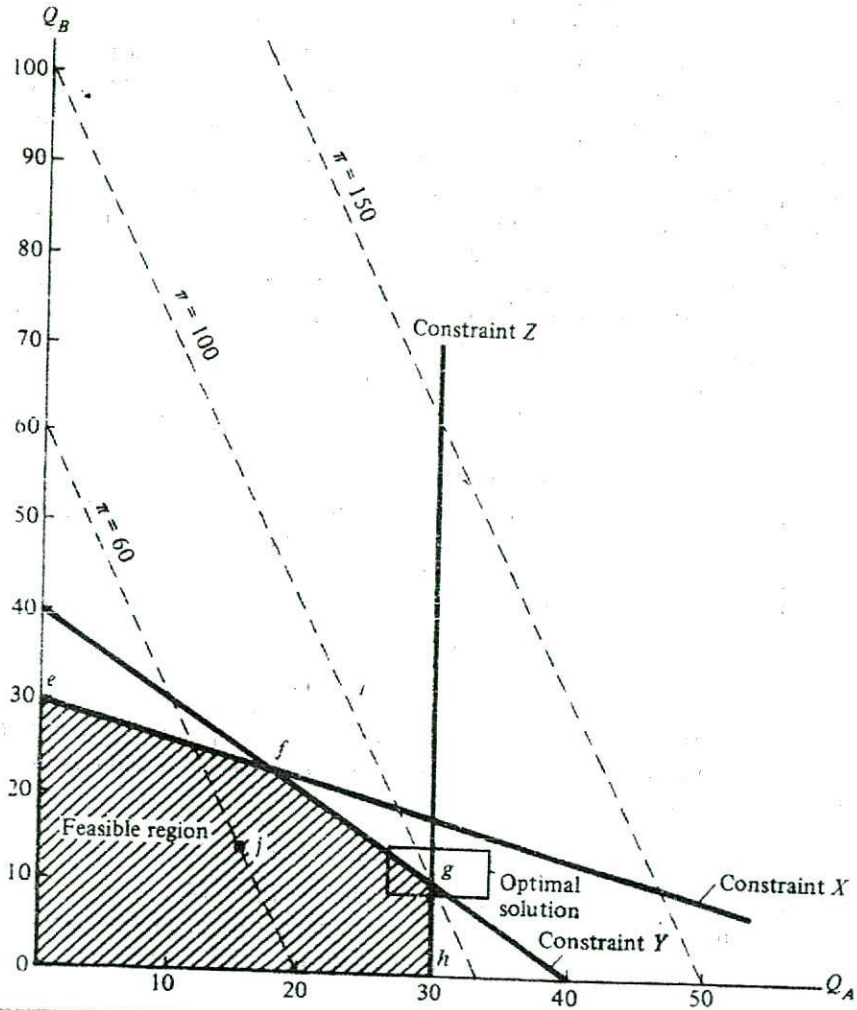


FIGURE 8.2 Solution to a Linear Programming Problem

TABLE 8.2 Hours of Time Allocated to Each Product

Machine	Hours of Time Allocated to Product A	Hours of Time Allocated to Product B	Hours Available	Slack Hours
X	30	30	90	30
Y	60	20	80	0
Z	60	0	60	0

timal point (g) in Figure 8.3. That point lies on constraints Y and Z but is below the constraint for machine X. This also shows that all of the available hours on machines Y and Z are used but that there are unused hours on machine X.

The concept of binding and nonbinding constraints has implications for determining the opportunity cost of the limited resources. With regard to the specific problem, the available hours of time on machines Y and Z have an opportunity cost; if one or more of these hours were reallocated to some other use, total profit would be reduced. That is, reducing hours available on machine Y or Z would result in a reduction in the rate of one or both outputs Q_A and Q_B. The result of this reduced output would be a smaller profit. But there are 30 hours of time available on machine X that could be allocated to some other use without affecting profit. Thus those hours have zero opportunity cost as far as producing products A and B are concerned.

Also note in Figure 8.3 that the profit-maximizing solution occurs at a corner of the feasible region. Because both the objective function and the constraints are linear, the optimal solution will always occur at one of the corners.[2] By having to evaluate only the corner solutions, the number of computations is greatly reduced. For instance, in the example, points e, f, g, and h are the corners and therefore are the only points that need to be considered. In the computer programs written to solve linear programming problems, the approach is to pick one corner arbitrarily, evaluate the objective function at that point, and then systematically move to other corners that offer higher profits until no other corner yielding greater profit is found. This process greatly reduces the time necessary to determine the optimal solution.

In Figure 8.4, two different isoprofit lines, π_1, and π_2, are drawn. The optimal solution for the original profit equation,

$$\pi_1 = 3Q_A + Q_B$$

(labeled π_1 in the figure), has been shown to be at point g. For the other objective function, there is one dollar of profit for each unit of Q_A and two dollars of profit per unit of Q_B. Thus the profit equation is

$$\pi_2 = Q_A + 2Q_B$$

For profit function π_2, the optimal solution is at corner point f. The solution to any linear programming problem will always be at a corner point, and the slope of the objective function is critical in determining which corner point is optimal.

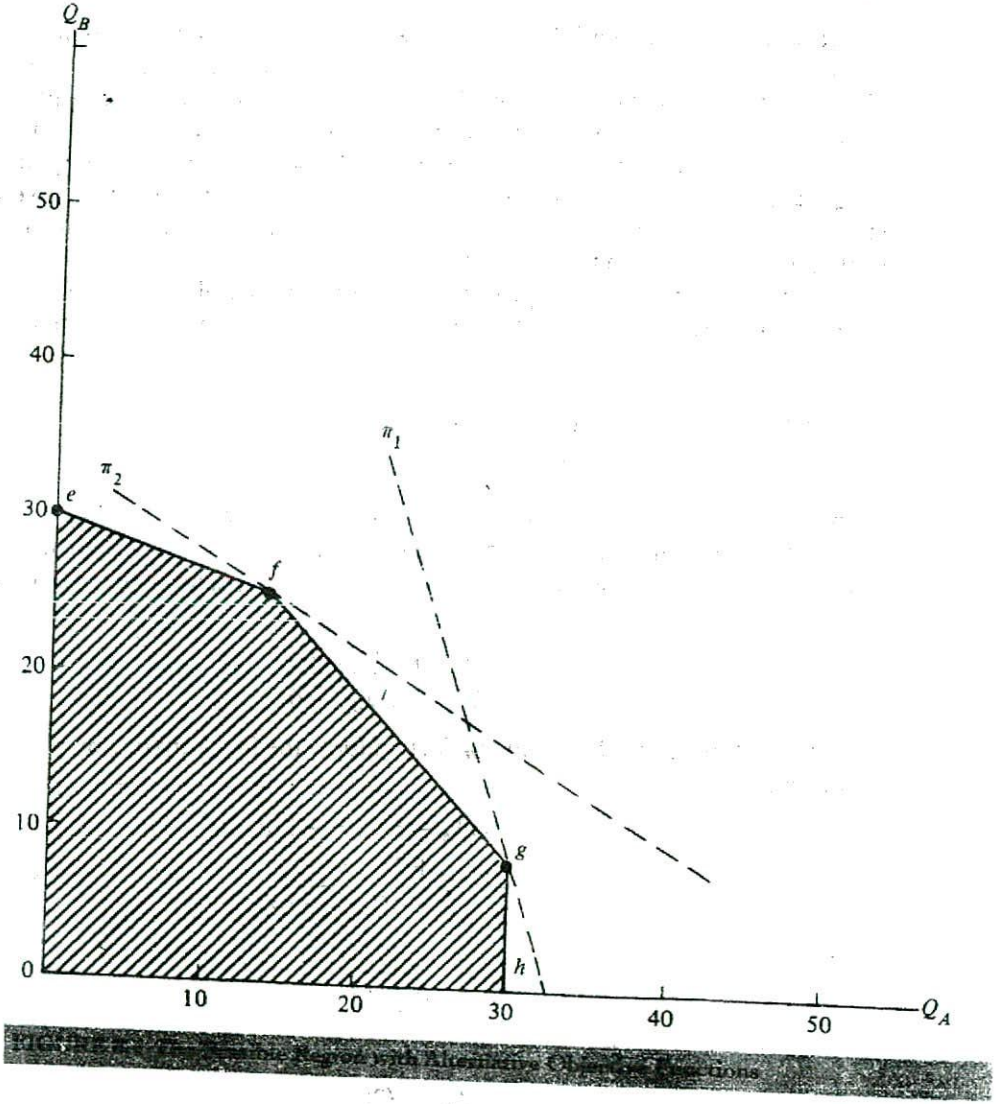
Algebraic Solution

The graphic approach just used demonstrates the conceptual solution to linear programming problems and actually could be used to solve some problems. However, if there are more than two decision variables, the graphic approach cannot be used. Fortunately, there is an algebraic technique that, at least conceptually, can be used to solve linear programs of almost any size.

The problem considered in the preceding section involved an objective function, three inequalities denoting the resource constraints, and the nonnegativity requirements. That is,

---

[2]It is possible that the highest isoprofit line will coincide with one of the sides of the feasible region, such as fg in Figure 8.4. In that case, points f, g, and all others in between are optimal in that they yield the same level of profit. Still a corner solution, f or g, will optimize the objective function. The possibility of multiple solutions is discussed in detail later in the chapter.



maximize: $\pi = 3Q_A + 1Q_B$
subject to: $1Q_A + 3Q_B \leq 90$
$2Q_A + 2Q_B \leq 80$
$2Q_A \leq 60$
$Q_A, Q_B \geq 0$

As outlined earlier in this chapter, the general approach to solving the problem is to identify corners of the feasible region and then evaluate profits at each corner. Because the corners are defined by the intersection of the constraints, those corners can be found by identifying which constraints combine to form that corner and then solving those two constraint equations for the values of the decision variables Q_A and Q_B at that point of intersection.

Refer again to the feasible region shown in Figure 8.4. The only points that have to be evaluated are the corners *e, f, g,* and *h*. Technically, the origin, 0, is a corner of the feasible region, but it is obvious that profits cannot be maximized at that point. Also, point *h* can be dismissed as a possibility because point *g* yields the same output rate for *A* and a greater output rate for *B;* thus *g* is unequivocally better. Therefore, only points *e, f,* and *g* require consideration. Hence the linear programming problem can be solved by finding the values (Q_A, Q_B) at points *e, f,* and *g;* evaluating the profit rate at each; and selecting that combination that yields the greatest profit.

Point *e* is the intersection of constraint *X* with the vertical (Q_B) axis. Thus Q is determined by substituting $Q_A = 0$ (the equation for the vertical axis) into the first constraint

$$0 + 3Q_B = 90$$

and solving for Q_B, yielding

$$Q_B = 30$$

Thus, at point *e,* the values of the decision variables are $Q_A = 0$ and $Q_B = 30$.

Point *f* is the intersection of constraints *X* and *Y.* By writing these constraints as equalities and solving them simultaneously, the value of Q_A and Q_B can be determined. The two equations are

$$Q_A + 3Q_B = 90$$
$$2Q_A + 2Q_B = 80$$

To solve for the values of Q_A and Q_B, first multiply the first equation by -2 and add the two equations:

$$\begin{aligned} -2Q_A - 6Q_B &= -180 \\ 2Q_A + 2Q_B &= 80 \\ \hline -4Q_B &= -100 \end{aligned}$$

Thus

$$Q_B = 25$$

Now, substitute $Q_B = 25$ into the first equation to find the value of Q_A. That is,

$$Q_A + 3(25) = 90$$

or

$$Q_A = 15$$

Thus, at point *f,* the values of the decision variables are $Q_A = 15$ and $Q_B = 25$.

Finally, point *g* is the intersection of constraints *Y* and *Z*. The values of Q_A and Q_B are found by simultaneously solving the equations

$$2Q_A + 2Q_B = 80$$
$$2Q_A = 60$$

which yields the solution $Q_A = 30$ and $Q_B = 10$.

Now, by substituting the values (Q_A, Q_B) at each corner into the profit function

TABLE 8.3 Values of the Decision Variables and Profit at Each Corner Point

	Decision Variables		
Corner Point	Q_A	Q_B	*Profit*
0	0	0	0
e	0	30	30
f	15	25	70
g	30	10	100
h	30	0	90

$$\pi = 3Q_A + Q_B$$

the profit at each corner is determined. For example, at point *f*, $Q_A = 15$ and $Q_B = 25$. Therefore,

$$\pi = 3(15) + 1(25) = 70$$

The values of the decision variables and the associated profit rate at each corner are reported in Table 8.3. The values at the origin and point *h* have been included, even though it is known that they cannot be profit-maximizing points.

Corner point *g* yields the highest profit ($\pi = 100$), and therefore the optimal solution for the firm is to produce 30 units of Q_A and 10 units of Q_B. Of course, this is the same solution obtained using the graphic approach.

As the linear programming problem becomes larger, the number of computations increases rapidly. For this reason, most solutions are found by using computers. It is not unusual for a problem to have 10, 20, or more decision variables and literally dozens of constraints. It simply is not practical to even attempt to solve such problems without the aid of a computer. Numerous computer programs for solving linear programming problems are available and are easily used, including the TOOLS software package.

Key Concepts

- The graphic approach to solving a linear programming problem consists of graphing the feasible region (as defined by the system of constraints) and then shifting the objective function until an optimal solution is found at a corner of the feasible set.
- Solving a linear programming problem algebraically requires determining the values of the decision variables at each corner point of the feasible region and then evaluating the objective function for each set of the decision variable values so determined.
- The solution to a linear programming problem will always occur at a corner point of the feasible region. This greatly simplifies the number of computations required to find the optimal solution.

CONSTRAINED COST MINIMIZATION

Linear programming has been defined as an optimization technique for finding the set of values for the decision variables that maximizes or minimizes an objective function subject to constraints. In the preceding section, a profit function was maximized subject to resource constraints in the form of a limited number of available machine hours. In this section, the solution to a minimization problem is outlined. The approach is analogous to that used in the maximization problem.

Structuring the Problem

The example used here comes from agribusiness, where linear programming models have been widely used. Consider a milk producer whose objective is to feed the milk cows adequately but to do so at minimum cost. Suppose that an adequate feed ration consists of a minimum of 40 units of protein, 60 units of calcium, and 60 units of carbohydrates.

The manager must determine how much of two feeds, A and B, to use. One ton of feed A contains one unit of protein, three units of calcium, and one unit of carbohydrates. One ton of feed B contains one unit of protein, one unit of calcium, and six units of carbohydrates. Let the price of feed A be \$100 per ton and the price of feed B be \$200 per ton. These basic data are summarized in Table 8.4.

The problem is to find the quantities of the two feeds, X_A and X_B (the decision variables), to be purchased so that the feed cost (C) will be minimized. Thus, the problem is to minimize the objective function,

$$C = 100X_A + 200X_B$$

subject to the following minimum nutrition requirements or constraints:

$$\begin{aligned} 1X_A + 1X_B &\geq 40 \quad \text{(protein constraint)} \\ 3X_A + 1X_B &\geq 60 \quad \text{(calcium constraint)} \\ 1X_A + 6X_B &\geq 60 \quad \text{(carbohydrate constraint)} \end{aligned}$$

and the usual requirement that all decision variables be nonnegative:

$$X_A \geq 0 \qquad X_B \geq 0$$

TABLE 8.4 Summary of Data for the Cost-Minimization Problem

	Feed	
	A	***B***
Price per ton	\$100	\$200
	Units of Nutrients Per Ton of Feed	
	A	***B***
Protein	1	1
Calcium	3	1
Carbohydrates	1	6

Minimum Units Required Per Period
40
60
60

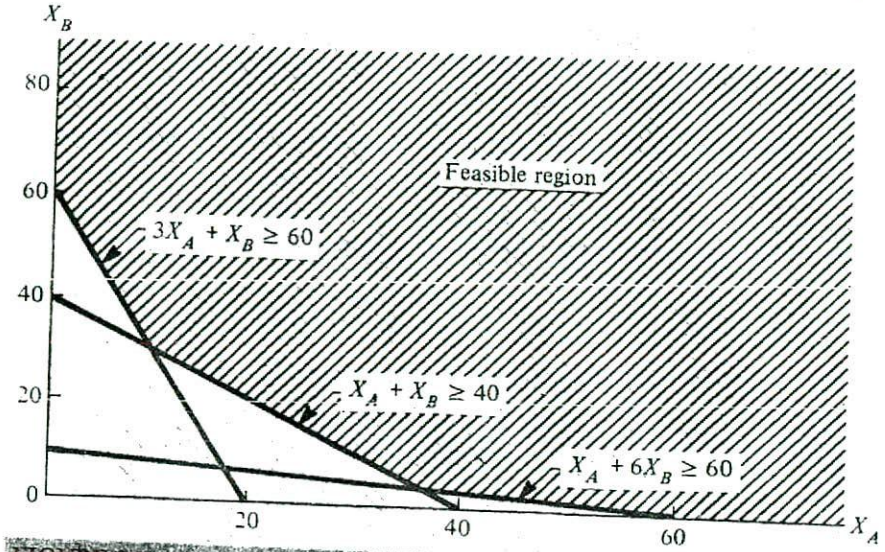


FIGURE 8.5 Constraints and Feasible Region for the Cost Minimization Problem

The constraints are plotted in a graph just as was done in the preceding problem. These constraints and the resultant feasible region are shown in Figure 8.5. Note that the feasible region extends upward and to the right of the constraints, in contrast to the preceding problem, where the feasible region extended to the left and below the resource constraints. This is due to the nature of the resource constraints. In the first problem, the resource constraints were all of the form "less than or equal to" (i.e., $\leq$), thus restricting feasible combinations of the decision variables to points on or below those constraints when shown graphically. In this problem the resource constraints are of the form "greater than or equal to," which restricts feasible combinations of the decision variables to points on or above those constraints.

To find the combination of feed inputs (X_A, X_B) that meets the nutrition requirements at minimum cost, think in terms of starting at the origin and making parallel shifts in the cost equation $C = 100X_A + 200X_B$ in a "northeasterly" direction until that cost function touches a point in the feasible region. The solution is shown in Figure 8.6. The cost function first touches the feasible set at point *g*, corresponding to 36 tons of *A* and 4 tons of *B*. This is the optimal solution in the sense that it is the least-cost of all combinations of the two feeds that meet the constraints. The cost of these feed inputs is \$4,400 [i.e., $C = 100(36) + 200(4)$]: there is no other combination of the two feeds (X_A, X_B) that satisfies the constraints and costs less than \$4,400.

Algebraic Solution

The algebraic approach is essentially the same as that used for the profit-maximization problem. First, find the values of X_A and X_B at each corner of the feasible region (i.e., at points *e*, *f*, *g*, and *h*) by solving the two intersecting constraint equations simultaneously at each point. For example, consider point *f* in Figure 8.6, which is the intersection of the first two constraints. Writing these as equalities gives

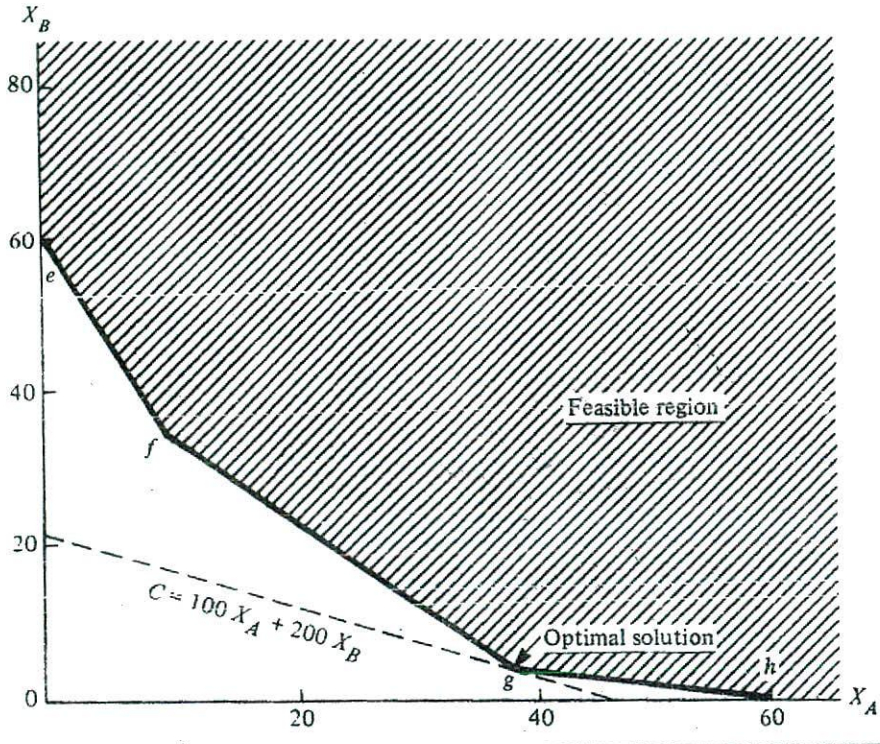


FIGURE 8.6 Solving the Cost-Minimization Problem

$$X_A + X_B = 40$$
$$3X_A + X_B = 60$$

Subtracting the second from the first yields

$$-2X_A = -20$$

or

$$X_A = 10$$

Now, substituting $X_A = 10$ into the first constraint and solving for X_B yields

$$10 + X_B = 40$$

or

$$X_B = 30$$

Thus, point f is ($Q_A = 10$, $Q_B = 30$). The next step is to evaluate the cost of that input combination by substituting X_A and X_B into the cost equation. Repeating this process at points e, g, and h yields the values of the decision variables and cost at each corner point. These are summarized in Table 8.5.

TABLE 8.5 Values of the Decision Variables and Cost at Each Corner Point

	Decision Variables		
Corner Point	X_A	X_B	*Cost*
e	0	60	$12,000
f	10	30	7,000
g	36	4	4,400
h	60	0	6,000

Clearly, the optimal solution is at point *g*. The manager should use 36 tons of feed *A* and 4 tons of feed *B*. There is no other input combination that can satisfy the constraints at a cost less than $4,400.

Whether the objective is minimizing or maximizing the objective function, the approach to solving the linear programming problem is essentially the same. First, use the constraints to determine the feasible region. Next, determine the values of the decision variables at each corner point of the feasible region. Then evaluate the objective function for each of those combinations, and select the combination that optimizes (i.e., minimizes or maximizes) that function.

Case Study

Linear Programming and Hospital Staffing

The allocation of a staff of nurses of differing specialties to the various wards and other departments is a major problem in most hospitals. A large institution may have hundreds of nurses, ranging from practical nurses to surgical specialists. Some wards need nurses on duty 24 hours per day, whereas others may need only one shift. Further complicating the problem, some nurses can work in several areas, whereas others can only be assigned to one. In addition, provision must be made for periods of vacation and substitutes for nurses who are absent because of illness. Most hospitals use temporary staff and overtime work to fill in where members of the regular nursing staff are unavailable. Generally, this alternative is more costly than using regular staff at standard wage rates.

The job of scheduling nurses could be a nightmare in a large hospital. However, the very nature of the problem makes it amenable to solution by means of linear programming. The objective function is to minimize the cost of providing nurses subject to constraints on the number and type of nurses required in each area of the hospital during each shift and the number of nurses available at each relevant period of time.

The administrator at one Chicago hospital faced these problems as she scheduled the assignments for more than 300 nurses. Historically, this job was done by hand and required more than three days of work each month. The acquisition of a personal computer and some linear programming software allowed her to complete this task in an afternoon. Not only did she save considerable time in the process, but the hospital was

able to significantly reduce the amount of overtime work and the hiring of temporary workers. Total wage costs were reduced by more than 10 percent, thus saving the hospital more than $1 million a year. ■

Sensitivity Analysis

Decision markers often are faced with questions of the what-if variety. For example, a production manager might be asked how the production process should be changed if the price of labor increased by 10 percent. That is, how should the mix of labor and capital be altered to adjust for the higher price of labor? Linear programming is an especially useful technique for answering questions of this kind. Given an original problem and its solution, it is a simple matter to change one or more of the parameters of that problem, solve it again, and then compare the original and new solutions. This is the essence of sensitivity analysis.

Consider the optimal feed ration problem in the previous section. Given the price of feed A of $100 per ton and the price of feed B of $200 per ton, it was determined that the combination of 36 tons of A and 4 tons of B is the least-cost combination (i.e., $4,400) of feeds that meets the nutrition constraints. Now, what if the price of feed A doubles to $200—what is the effect on least-cost input combination and its cost? Changing the price of feed A to $200 per ton and solving the problem again yields an optimal solution of 10 tons of A and 30 tons of B, with a total cost of $8,000. Doubling the price of feed A causes a substantial change in the optimal mix of feeds and a large increase in cost. Note that the amount of A used has declined from 36 tons to 10 tons, and the amount of B has increased. The firm has substituted B for A because the relative cost of A has increased.

Sensitivity analysis also can be used to measure the effect of a change in one of the constraints. For example, in the original problem, what if new research indicated that the animals needed only 50 units of carbohydrates per day rather than 60? Changing the right-hand side of the third constraint from 60 to 50 and solving the problem yields a cost-minimizing solution of 38 tons of A and 2 tons of B, and that combination has a total cost of $4,200. Thus, that change in the constraint resulted in a cost reduction of $200.

Finally, if a new feed that had three units of protein, five units of calcium, and seven units of carbohydrates were introduced at a price of $210 per ton, what would be the new optimal mix of the three feeds? Using the original data but adding a new decision variable for this feed (X_C) and changing the constraints appropriately yields the following problem:

$$\begin{aligned} \text{Minimize: } & C = 100X_A + 200X_B + 210X_C \\ \text{Subject to: } & 1X_A + 1X_B + 3X_C \geq 40 \\ & 3X_A + 1X_B + 5X_C \geq 60 \\ & 1X_A + 6X_B + 7X_C \geq 60 \\ & X_A, X_B, X_C \geq 0 \end{aligned}$$

It can be shown that the new solution is zero units of both A and B and 13.33 tons of the new feed (i.e., $X_A = 0$, $X_B = 0$, and $X_C = 13.33$), with a total cost of \$2,800. Thus, the new feed is used exclusively, and the total cost is reduced by \$1,600 compared to the original solution.

As indicated by these examples, sensitivity analysis involves solving a linear programming problem, then changing one or more component parts of the problem, solving the problem again, and comparing the two solutions. The coefficients of the objective function, the coefficients of the constraints, and/or the resource amounts on the right-hand side of the constraints can be changed. Also, a decision variable or a constraint could be added to or deleted from the problem.

Key Concepts

- The algebraic solution to a linear programming problem requires that the coordinates of each corner point of the feasible solution be determined and then the objective function evaluated for each set of coordinates.
- Sensitivity analysis involves solving a linear programming problem, changing one or more components of the objective function or the constraints, solving the new problem, and then comparing the solutions.

Example The Transportation Problem

One area where linear programming has been especially useful is in determining optimal shipping patterns. A typical problem is to determine the minimum cost for shipping output from manufacturing plants to dealers. In general, the problem is to determine how much to ship from each source or supply point to each destination or demand point, such that the total shipping costs are minimized. The constraints are the maximum production rates at each source and the quantity demanded at each destination.

Consider the following hypothetical example. An automobile manufacturer with plants in Detroit and Los Angeles must supply its dealers in Atlanta, Chicago, and Denver. Assign an index i to each plant (e.g., 1 for Detroit and 2 for Los Angeles) and an index j to each dealer (e.g., 1 for Atlanta, 2 for Chicago, and 3 for Denver). The data in the following table show the transportation cost per car from each plant to each dealer (i.e., C_{ij}), the maximum production per period at each plant, and the number of cars demanded at each dealership.

	Transportation Cost per Car			
Plant	***Atlanta (1)***	***Chicago (2)***	***Denver (3)***	***Number of Cars Produced (supply)***
Detroit (1)	200 (C_{11})	100 (C_{12})	300 (C_{13})	3,000
Los Angeles (2)	400 (C_{21})	300 (C_{22})	200 (C_{23})	5,000
Number of cars demanded	3,000	4,000	1,000	

There are six decision variables, X_{ij} (where $i = 1, 2$, and $j = 1, 2, 3$), representing the number of cars shipped from each plant to each dealer. For example, X_{23} would refer to shipments from Los Angeles to Denver. The objective is to determine the number of

cars to be shipped from each plant to each dealer that will minimize total shipping costs. That is, minimize

$$C = \sum_i \sum_j C_{ij} X_{ij}$$

The constraints are (1) the total number of cars shipped from each plant (e.g., $X_{11} + X_{12} + X_{13}$ would be total shipments from Detroit) must be equal to or less than the maximum output rate for that plant; and (2) shipments to each dealer (e.g., $X_{11} + X_{21}$ would be the number of cars shipped to Atlanta) must be at least as great as the quantity demanded.

Thus, the linear program is to find those values of the decision variables X_{ij} that will minimize

$$C = 200X_{11} + 100X_{12} + 300X_{13} + 400X_{21} + 300X_{22} + 200X_{23}$$

subject to the production or supply constraints

$$X_{11} + X_{12} + X_{13} \leq 3{,}000$$
$$X_{21} + X_{22} + X_{23} \leq 5{,}000$$

the demand constraints

$$X_{11} + X_{21} \geq 3{,}000$$
$$X_{12} + X_{22} \geq 4{,}000$$
$$X_{13} + X_{23} \geq 1{,}000$$

and the nonnegativity constraints

$$X_{ij} \geq 0 \qquad i = 1, 2; \qquad j = 1, 2, 3$$

Clearly, this problem cannot be solved graphically, and the algebraic approach would be cumbersome. However, the problem can be solved easily using a microcomputer to yield the following optimal values of the decision variables:

	Least-Cost Shipment Pattern to			
From	***Atlanta***	***Chicago***	***Denver***	***Total Production***
Detroit	$X_{11} = 3{,}000$	$X_{12} = 0$	$X_{13} = 0$	3,000
Los Angeles	$X_{21} = 0$	$X_{22} = 4{,}000$	$X_{23} = 1{,}000$	5,000
Total demand	3,000	4,000	1,000	

Note that the supply (production) and demand constraints are met. For example, total production is 3,000 in Detroit and 5,000 in Los Angeles, which are the maximum output rates in those plants. Further, the dealers in each city receive exactly the number of cars necessary to meet demand. The total transportation cost is $2,000,000. No other set of values for the decision variables would meet the constraints and result in lower transportation costs.

The transportation problem can become very large. If there are m supply sources and n demand points there will be $m \times n$ decision variables, $m + n$ resource constraints, and $m \times n$ nonnegativity constraints. For example, for a problem involving an automobile manufacturer with 10 plants and 200 dealers there would be 2,000 decision variables, 10 supply constraints, 200 demand constraints, and 2,000 nonnegativity constraints!

SPECIAL PROBLEMS IN LINEAR PROGRAMMING

There are several special situations that may be encountered in linear programming problems. Some do not pose a problem but can generate rather curious results. In other cases, the result is that there is no optimal solution to the problem.

Multiple Solutions

If the objective function has the same slope as one of the constraints, the result will be an infinite number of optimal combinations of the decision variables. Consider the following problem:

$$\begin{aligned} \text{maximize:} \quad & \pi = 10X_1 + 5X_2 \\ \text{subject to:} \quad & X_1 + 2X_2 \leq 60 \quad \text{(constraint } A\text{)} \\ & 2X_1 + X_2 \leq 60 \quad \text{(constraint } B\text{)} \\ & X_1 \leq 27 \quad \text{(constraint } C\text{)} \\ & X_1 \geq 0, X_2 \geq 0 \end{aligned}$$

In this problem, the objective function and constraint B both have a slope equal to -2. This can be seen by rewriting each of those relations as a function of X_2. That is,

$$\begin{aligned} \textit{PROFIT FUNCTION:} \quad & X_2 = \frac{\pi}{5} - 2X_1 \\ \textit{CONSTRAINT } B\textit{:} \quad & X_2 \leq 60 - 2X_1 \end{aligned}$$

The problem is shown graphically in Figure 8.7a.

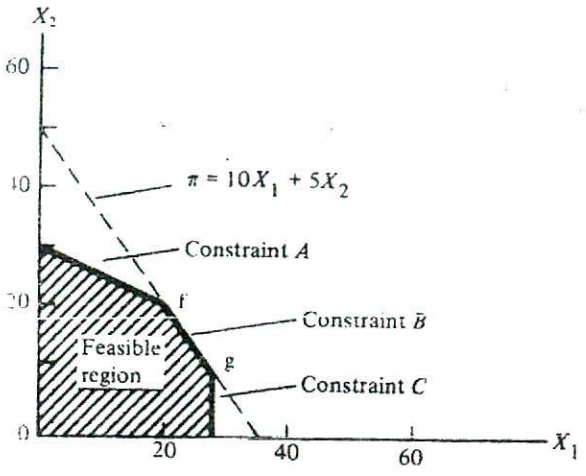
The optimal level of the objective function is coincident with constraint B between points f and g. Thus all combinations of (X_1, X_2) in that interval yield the same profit, and therefore all are equally profitable. This situation does not pose a problem, and the solution is exactly as outlined earlier. The corner solutions are evaluated and both f and g yield the same profit. Thus, either can be used as the optimal solution.

Redundant Constraints

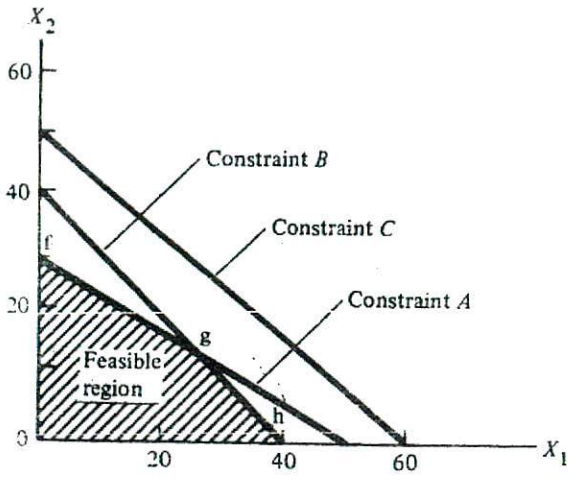
Sometimes, one or more constraints is unnecessary or redundant. Consider the following constraint set:

$$\begin{aligned} A\text{:} \quad & X_1 + 2X_2 \leq 50 \quad \text{(constraint } A\text{)} \\ B\text{:} \quad & X_1 + X_2 \leq 40 \quad \text{(constraint } B\text{)} \\ C\text{:} \quad & 5X_1 + 6X_2 \leq 300 \quad \text{(constraint } C\text{)} \\ & X_1 \geq 0, X_2 \geq 0 \end{aligned}$$

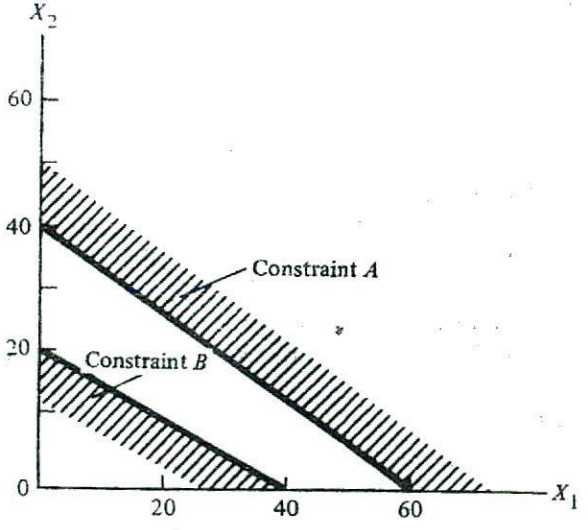
The feasible region for these constraints is shown in Figure 8.7b. Note that this region is defined by constraints A, B, and the nonnegativity requirements. Constraint C is redundant because if either constraint A or B is satisfied, so is constraint C. This does not present a problem in solving the problem. It only means that the information contained in constraint C is irrelevant and therefore unnecessary. Because constraint C is nonbinding, the resource associated with that constraint has a zero opportunity cost. The solution would proceed as before by evaluating the objective function at corner points f, g, and h.



(a) Multiple solutions



(b) Redundant constraint



(c) No feasible solution

FIGURE 8.7 Special Problems with Linear Programming

No Feasible Solution

A serious problem arises when the set of constraints is such that there are no values of the decision variables that simultaneously satisfy all the constraints. The following constraints have been graphed in Figure 8.7c:

$$2X_1 + 3X_2 \geq 120 \quad \text{(constraint } A\text{)}$$
$$X_1 + 2X_2 \leq 40 \quad \text{(constraint } B\text{)}$$

Only points below $X_1 + 2X_2 = 40$ satisfy constraint *B*, and only points above $2X_1 + 3X_2 = 120$ satisfy constraint *A*. But there are no points that satisfy both these conditions simultaneously. Therefore, there is no feasible solution.

In a simple problem such as the example shown, the lack of a feasible solution is easily seen. However, in large linear programming problems with many constraints, there may be no way of knowing that there is no solution until the computer program is run. In such cases, obtaining a feasible solution means that one or more constraints must be modified. For example, additional hours of machine time may have to be acquired in order for any production to take place.

Key Concepts

- In some linear programming problems, multiple solutions and redundant constraints occur. Neither of these possibilities poses a problem for solving the linear program.
- Sometimes the constraints define a region for which there is no feasible solution. In this case, obtaining a solution (i.e., finding nonzero values of the decision variables that meet all the constraints) will require that one or more constraints be changed by augmenting the resource associated with those constraints.

THE DUAL PROBLEM

It is known that when a resource constraint is binding, there is an opportunity cost associated with that resource. Consider a profit-maximization problem where the resource constraints are hours of time available on machines *X*, *Y*, and *Z*. Assume that the constraints or hours of time on machines *Y* and *Z* are binding. If one hour of time on either of those machines had to be given up, profit would fall. But if the constraint for hours of machine *X* is not binding, this means that there is a zero opportunity cost for this resource.

In terms of planning for expansion of facilities, it would be useful to have an estimate of the value or opportunity cost of time on machines *Y* and *Z*. For example, if the firm could hire one more hour of machine time, should it hire more hours of *Y* or *Z*? Suppose it was known that the opportunity cost of one hour of time on machines *Y* and *Z* are \$20 and \$10, respectively. This means that an additional hour on machine *Y* would add \$20 to profit in contrast to adding only \$10 to profit for an additional hour on machine *Z*. The dual linear program is a way to determine the opportunity cost of the resource associated with each constraint.

For every linear programming problem (called the *primal problem*), there is an associated problem, referred to as the *dual problem*. If the objective of the primal problem is maximization of an objective function, the objective of the dual problem is minimization of an associated objective function and vice versa. For example, if the objective of the primal problem is to maximize production subject to a set of machine-hour constraints, the objective of the dual problem would be to minimize the cost of the resources (i.e., the machine hours) subject to a constraint on production. The dual and primal problems are said to be symmetric because a solution to one is a solution to both. For example, in the production problem used at the beginning of the chapter, the objective was to find the values Q_A and Q_B that maximized profit. Determining

this combination (Q_A, Q_B) also determined the hours of time used on each machine. The dual problem would find the set of hours on each machine that minimized the cost of the resources devoted to the production of the outputs Q_A and Q_B. The solution to the dual problem would allocate the same number of machine-hours to each product as did the primal problem.

Structuring the Dual Problem

The interpretation of the dual will be made clearer by setting up, solving, and interpreting a dual problem. Recall the original profit-maximization problem on page 274. This will be referred to as the primal problem:

$$\begin{array}{ll} \text{maximize: } \pi = & 3Q_A + 1Q_B \\ \text{subject to:} & 1Q_A + 3Q_B \leq 90 \\ & 2Q_A + 2Q_B \leq 80 \\ & 2Q_A + 0Q_B \leq 60 \end{array}$$

The dual problem is constructed by using the columns of the coefficients of the primal problem as shown by the dashed lines. Define three new variables C_x, C_y, and C_z that represent the opportunity cost of hours on machines X, Y, and Z, respectively. The objective function for the dual problem is to

$$\text{minimize: } C = 90C_x + 80C_y + 60C_z$$

That is, the objective is to minimize the opportunity cost of the firm's scarce resources (i.e., hours of machine time).

The coefficients of first constraint in the dual problem are the coefficients of Q_A from the first column of the primal problem. That is,

$$1C_x + 2C_y + 2C_z \geq 3$$

meaning that the value of machine-hours used to produce one unit of product A must be at least as great as the profit on that unit of output. Thus one unit of output A requires one hour of time on machine X, two hours on Y, and two hours on Z, and those hours are valued at C_x, C_y, and C_z, respectively.

The second constraint is based on the coefficients in the second column in the primal problem. Hence

$$3C_x + 2C_y + 0C_z \geq 1$$

Note that output B does not require processing on machine z, so that there is no cost associated with that machine.

As with the primal problem, there are nonnegativity requirements for the dual, that is,

$$C_x, C_y, C_z \geq 0$$

These requirements constrain opportunity costs from being negative. A negative opportunity cost would make no sense because it would mean that profit could be increased by giving up the resource.

Finally, as there are three decision variables but only two resource constraints, there will be at most two decision variables with nonzero values.

Solving the Dual Problem

The solution to the dual problem follows exactly the same steps as before. That is, define the feasible region, identify the values of the decision variables at each corner, evaluate the objective function using each set of values of the C_i, and select that set that yields the minimum value for the objective function.

Because there are three variables, a graphic solution would require three-dimensional analysis and would be extremely difficult to draw and interpret. An algebraic approach is more straightforward, but with three constraints and three decision variables, even this can be cumbersome. The solution to this particular problem is simplified because it is known that constraint X in the original problem is not binding, and therefore the opportunity cost of machine X hours is known to be zero (i.e., $C_x = 0$). Thus the problem reduces to

$$\begin{aligned} \text{minimize: } & C = 80C_y + 60C_z \\ \text{subject to: } & 2C_y + 2C_z \geq 3 \\ & 2C_y \geq 1 \\ & C_y, C_z \geq 0 \end{aligned}$$

Further, it is known that both constraints Y and Z are binding so that both C_y and C_z must be positive. As this can only occur at the intersection of the two constraints (not at the corners on the horizontal and vertical axes), consider the constraints as equalities and solve them simultaneously. That is,

$$\begin{aligned} 2C_y + 2C_z &= 3 \\ 2C_y &= 1 \end{aligned}$$

Subtracting the second equation from the first yields

$$2C_z = 2$$

or

$$C_z = 1$$

and it follows that

$$C_y = 0.5$$

These values, $C_y = 0.5$ and $C_z = 1$, represent the opportunity cost per hour of time on machines Y and Z, respectively. That is, a one-hour reduction in time available on machine Y would reduce profit by \$0.50 and on machine Z by \$1.00.

Evaluating these costs using the objective function yields a cost of \$100, that is,

$$\begin{aligned} C &= 90(0) + 80(0.5) + 60(1) \\ &= \$100 \end{aligned}$$

This minimum cost is the same as the maximum rate of profit for the primal problem. That is, at the point of optimal resource allocation, profit equals the value or the opportunity cost of the resources being used to generate that profit.

Thus the linear programming analysis has established the opportunity cost of an hour of time on each machine as $C_x = 0$, $C_y = 0.50$, and $C_z = 1$. The value of each C_i is the increase in profit associated with one additional hour of time on the ith machine.

The opportunity cost of any resource is its value to the firm. The term *shadow price* is often used in the context of dual programming problems to describe this opportunity cost. That is, the C_i in the problem are the shadow prices of the machine-hours. As the firm makes plans to add more machine capacity, it can compare the shadow price to the market or acquisition price of additional hours of machine time. If the shadow price exceeds the acquisition price, profit can be increased by acquiring more hours. Thus, the dual problem provides direction to management when making decisions about expanding productive capacity.

Case Study

Cost Minimization at Wellborn Cabinet, Inc.

Wellborn Cabinet, Inc. is an integrated producer of cabinets located in Ashland, Alabama. The cabinet industry is very competitive with numerous small producers located throughout the United States, and the entry of foreign producers is making conditions even more competitive. Unlike most of its competitors, Wellborn's operation includes a sawmill and drying kilns (for treating green wood) so that it can make its own lumber. Each year Wellborn purchases about $1.4 million in wood materials, including various types, sizes, and grades of logs (e.g., hardwood and common) and lumber (e.g., dry and green). The firm faced two questions: (1) Could raw materials costs be reduced while still maintaining the same rate of output of cabinets? and (2) Was the capacity of the sawmill and drying kilns adequate given that the firm's output would be increasing (i.e., did these two resources have a positive shadow price)?

The firm did not have the expertise to make the necessary analysis, so management sought help from faculty members in the School of Forestry at Auburn University. The professors determined that both questions could be answered using linear programming. That is, the optimal mix of logs and lumber could be determined by the primal problem and the shadow price of the sawmill and drying kilns computed by solving the dual problem.

The objective was to minimize the cost of procuring raw materials, and the objective function had 116 decision variables, including four types of lumber (i.e., grades 1 and 2 and types dry and green) and 112 different types of logs. The logs were classified by grade (i.e., hardwood and common), length (8, 10, 12, and 14 feet), and diameter (9 to 22 inches). There were 119 resource constraints in the model, as outlined here:

Resource	*Number of Constraints*
Sawmill capacity	1
Drying capacity	1
Wood requirements for cabinets	1
Supply of each type of log	112
Supply of each type of lumber	4

In addition, there were 116 nonnegativity requirements, one for each decision variable.

The solution of this linear programming problem indicated that total raw material cost would be minimized by meeting 88 percent of its wood requirements by buying only number 2 grade logs with a diameter of 9 to 15 inches and running these through the sawmill and drying kilns. The remaining 12 percent of its needs should be met by buying number 2 common green lumber. It was estimated that the company would save about $412,000 annually in the cost of raw materials, a 32 percent reduction from current levels. Also, a positive shadow price on the drying kiln constraint indicated that this facility was currently being used at capacity (i.e., there were no slack hours), and that it would have to be expanded if the output of cabinets were increased.

Some time after the study was completed, Paul Wellborn, the firm's president, wrote to the university: "We cannot follow the guidelines set forth by the model analysis 100 percent, but we are following it as closely as possible. We feel that we can look for a savings of up to $100,000 in this calendar year on solid wood raw materials purchases." ■

SOURCE: H. Carino and C. LeNoir, Jr., "Optimizing Wood Procurement in Cabinet Manufacturing." *Interfaces* 18(March–April 1988): 2.

Key Concepts

- For every linear programming problem, there exists an associated linear programming problem referred to as the dual problem.
- If the primal requires maximizing an objective function, the dual problem will involve minimizing an objective function.
- The solution to a dual problem results in estimates of the opportunity cost or shadow price of the resources that constrain the primal problem.

SUMMARY

Linear programming is a technique for solving constrained optimization problems where the objective function and the resource constraints are linear. Although the early applications of this tool were in the production area, linear programming has been successfully used in marketing, finance, transportation, and most other functional areas of management. The assumption of linear relationships implies that there are constant returns to scale in production and that output and input prices are constant. These relationships imply that cost and profit per unit are constant for all levels of output.

Conceptually, linear programming problems can be solved either graphically or algebraically. The graphic approach consists of identifying the feasible region and then shifting the objective function until an optimal solution is found at a corner of that feasible set. In the algebraic approach, the corners of the feasible set are identified and the values of the decision variables determined at those corners. Then the objective function is evaluated for each set of decision variables so identified and the maximum or minimum value selected. Sensitivity analysis involves solving a linear programming problem, then changing one or more parameters and comparing the original and new solutions.

If the objective function has the same slope as one of the constraints, multiple solutions to the linear program will occur. This poses no problem, as the standard solution techniques will still result in finding one of these several optimal solutions. In other cases, there is no feasible solution because the constraint system is such that there are no values of the decision variables that simultaneously satisfy all of the constraints. A resolution of this problem requires that one or more constraints be relaxed by augmenting the resources used in that constraint.

For every primal linear programming problem, there is an associated linear program referred to as the dual problem. If the primal problem requires maximizing an objective function, the dual linear program will be a minimization problem. The dual problem is structured using the columns of the primal problem. The solution to the dual problem results in estimates of the opportunity cost or shadow price of the resources that constrain the primal problem. The optimal solutions to the primal and dual problem yield the same value for the objective functions.

Discussion Questions

8-1. What assumptions about production, cost, and profit functions are implied in linear programming analysis?

8-2. In chapter 6, isoquant and isocost functions are used to demonstrate how to maximize output subject to a budget constraint. Describe the differences between that approach and the linear programming approach to solving a problem where the objective is to maximize output subject to one or more machine-time constraints.

8-3. Explain the relationship between the primal and dual linear programming problems. If the primal problem is to maximize production subject to machine-time constraints, how would the dual problem be stated?

8-4. Why is it that the resource associated with a binding constraint has a positive opportunity cost but the resource associated with a nonbinding constraint has a zero opportunity cost?

8-5. Why do only the corner points of the feasible region need to be evaluated in solving a linear program?

8-6. List two managerial problems (other than those described in the chapter) where linear programming tools could be used.

8-7. If a firm's production function is characterized by increasing returns to scale, what problem does this pose for using linear programming methods to determine optimal solutions to production problems?

8-8. Consider a situation where the linear programming problem is to maximize the profits associated with producing two products subject to a limited number of hours of time available on each of a number of machines. If a change in technology results in a reduction of processing time required on each machine, explain in general terms how the feasible region of production will change and how the optimal production rates of the two products will change.

8-9. How could the principles of linear programming be used at your college or university to allocate resources more efficiently?

Problems

8-1. Graph the region that is defined by each set of inequalities listed here.

a. $3x + 2y \leq 150$
$x + 2y \leq 80$
$x, y \geq 0$

b. $2x + 2y \leq 100$
$4x + 6y \leq 240$
$2x + 5y \geq 100$
$x, y \geq 0$

c. $10x + 5y \leq 50$
$2y \leq 15$
$3x \leq 9$
$x, y \geq 0$

d. $x + y \leq 40$
$2x + y \leq 60$
$3x + 3y \geq 60$
$x, y \geq 0$

8-2. Graph the feasible region defined by the following set of inequalities:

$$x + y \leq 40$$
$$2x + 4y \leq 100$$
$$3y \leq 60$$
$$x, y \geq 0$$

Using a graphic approach, determine that point in the feasible region (i.e., the values of x and y) that maximizes each of the following objective functions:

a. $z = x + 3y$
b. $z = 6x + 4y$

8-3. Given the following linear program,

$$\text{maximize } \pi = 4Q_A + 3Q_B$$

subject to the following machine-time constraints:

$$Q_A + 2Q_B \leq 100$$
$$2Q_A + Q_B \leq 80$$

and the nonnegativity constraints

$$Q_A, Q_B \geq 0$$

a. Solve the program using both an algebraic and graphic approach. Check to be sure that the optimal values of the decision variables are the same for both solutions.
b. Set up the associated dual problem and solve algebraically. Check to be sure that the value of the optimized objective function is the same for both the primal and dual problems.
c. What is the opportunity cost of one hour of time on each of the machines?

8-4. The officer in charge of a military mess hall has been ordered to design a minimum-cost survival-type meal that could be used in the event of a serious emergency. The meal is to consist only of milk and ground beef but must provide the following nutrient units: calories—300, protein—250, and vitamins—100. The nutrient content per ounce of each food is as follows:

	Milk	*Ground Beef*	*Minimum Units*
Calories	20	15	300
Protein	10	25	250
Vitamins	10	4	100

Milk can be purchased at $0.02 per ounce, and the price of ground beef is $0.04 per ounce.

a. Use linear programming to determine the composition of the lowest-cost meal (i.e., ounces of milk and beef) and the cost of that meal.
b. Set up the associated dual problem and explain how the shadow price (i.e., opportunity cost) of calories, protein, and vitamins would be determined.

8-5. National Publishing produces textbooks in plants in Boston, Atlanta, St. Louis, Denver, and San Francisco, which are then shipped to distribution centers in Newark, Chicago, Dallas, and Los Angeles. National is publishing a new managerial economics text and must supply its distribution facilities. The relevant data on quantity demanded at each distribution center, production capacity at each plant, and cost of shipping a book from each plant to each distribution center are shown here:

	Distribution Center				
Manufacturing Plant	***Newark***	***Chicago***	***Dallas***	***Los Angeles***	***Production Capacity***
	Shipping Costs per Unit				
1. Boston	$0.20	$0.35	$0.40	$0.60	40,000
2. Atlanta	0.35	0.40	0.45	0.50	10,000
3. St. Louis	0.30	0.20	0.30	0.40	15,000
4. Denver	0.50	0.40	0.30	0.30	15,000
5. San Francisco	0.70	0.50	0.45	0.20	20,000
Demand	20,000	40,000	30,000	10,000	

The president of National wants to know how to supply each distribution center to minimize the total shipping costs of meeting the demands at each center. Set up the transportation linear program to solve this problem. How many decision variables are there? What is the largest number of decision variables that can be expected to be nonzero?

8-6. The economics department at Southern State University produces two products—teaching, measured in student credit hours taught (H), and research, measured in pages published in professional journals (P). In any academic term, the department has 8,250 faculty hours to devote to teaching and research activities. Output is valued at $40 per credit hour and $1,000 per page published in journals. It is estimated that it takes 2.2 faculty hours per quarter to produce one student credit hour and 24 faculty hours per page published in a journal. In order to meet its mandate from the state legislature, the department must generate at least 1,800 student credit hours per quarter; to maintain credibility in the economics profession, the department must publish at least 120 pages of research output each quarter.

a. Set up the linear programming problem and draw a graph of the feasible region.
b. Solve the problem algebraically to determine how the department head should decide the output mix between credit hours and pages in order to maximize the value of departmental output.

8-7. Unique Software, Inc. produces two different video games, Firedarter and Paramedic, for the children's market. Each Firedarter game produced requires 0.2 hours of inspection time, 0.1 hours of packaging time, and 2.0 hours of assembly time. Each unit of the Paramedic program requires 0.1 hours of inspection, 0.2

hours of packaging, and 2.4 hours of assembly. The profit per unit is $4 on Firedarter and $6 on Paramedic. There are 200 hours of inspection time available, 300 hours of packaging time, and 2,400 hours of assembly time.

a. Solve this problem graphically.
b. Solve this problem algebraically.
c. What is the shadow price of assembly time? Explain.

8-8. Northwestern, Inc., a profit-maximizing firm, publishes textbooks using secretaries, editors, typesetters, and bindery workers. Given current staffing levels, a linear programming analysis has estimated shadow prices for each type of labor. The annual wage rates for these workers also are shown.

	Shadow Price	*Annual Wage*
Secretaries	$25,000	$20,000
Editors	48,000	35,000
Typesetters	20,000	25,000
Bindery	-0-	18,000

a. Given its current budget, should Northwestern change its mix of workers? Explain.
b. If Northwestern could add one worker, what type should it be? Why?

8-9. Set up the following linear programming problem graphically:

Max: $Z = 5Q_x + 10Q_y$

subject to:

$$\begin{aligned} Q_x + Q_y &\leq 5 \\ 2Q_x + 4Q_y &\leq 20 \\ Q_x + 4Q_x &\leq 4 \\ 2Q_x + 4Q_y &\leq 4 \\ Q_x, Q_y &\geq 0 \end{aligned}$$

a. Does constraint (2) affect the solution? Explain.
b. Can a solution be found if the following constraint is added? Why or why not?

$$4Q_x + 8Q_y \geq 48$$

8-10. A firm has m plants, each with a maximum supply capacity per period, and n warehouses, each having a demand requirement. The shipping cost is c_{ij} per unit shipped from plant i to warehouse j.

a. How many terms will be in the objective function?
b. How many total constraints will there be?
c. Set up the general problem of minimizing total transportation costs subject to meeting the supply-and-demand constraints.
d. What is the maximum number of decision variables that can be expected to be nonzero?

Computer Problems

The following problems can be solved by using the TOOLS program (downloadable from www.prenhall.com/petersen) or by using other computer software.

8-11. The tax commission of a state government employs 150 CPAs, 250 bookkeepers, and 40 investigators to audit state income tax returns. All employees work 2,000 hours per year. The number of hours required of each type of labor to audit different types of tax returns and the average additional tax revenue collected as a result of the audit are as follows:

Type of Return	*Required Time (hours) for:*			*Additional Tax Revenue Collected per Return Audited*
	CPA	*Bookkeeper*	*Investigator*	
Individual	2	4	3	$ 350
Small business	4	7	10	900
Corporation	30	15	24	2,400

a. Set up the linear programming problem to maximize the amount of additional tax revenue collected subject to constraints on time available by each type of worker.
b. Solve this problem to determine the revenue-maximizing number of audits of each type of return.
c. Set up, solve, and interpret the dual problem.
d. If the agency were faced with a cut in its budget and had to reduce its work force, which type of worker (i.e., CPA, bookkeeper, or investigator) should be the first to be terminated? Explain.
e. What if the state legislature provided additional funds to the tax commission, but specifically directed that they be used to hire five additional CPAs? How many additional audits could be performed? Is this decision by the legislature consistent with economic efficiency? Explain.

8-12. The production of type *A*, *B*, and *C* transistors requires processing in each of five areas of a firm's manufacturing facility. The profit per unit on these products is $0.07, $0.06, and $0.08, respectively. The processing times required in each area (in minutes) and the total minutes available per production period are shown here.

Transistor	*Time Required in Area (minutes)*				
	I	*II*	*III*	*IV*	*V*
A	3	1	2	5	2
B	1	1	4	3	5
C	2	2	1	4	4
Total time available in area (in minutes)	3,500	2,000	3,000	5,000	3,000

a. Determine the profit-maximizing production rates for the three products. What is the maximum profit?
b. Determine the shadow price (i.e., opportunity cost) of 1 minute of time in each of the production areas of the plant.

8-13. Skoshi Motors, Inc. has built two plants and three regional dealers in the United States and Canada. Use the following information to determine shipments from each plant to each dealer and the production rate at each plant that will minimize transportation costs. What is the minimum cost?

Plant	*Maximum Production Rate*
Cincinnati	90,000
Dallas	240,000

Dealer	*Number of Cars Required*
Seattle	70,000
Los Angeles	120,000
New York	140,000

From: / *To:*	*Shipping Cost per Car*		
	Seattle	*Los Angeles*	*New York*
Cincinnati	310	380	190
Dallas	260	190	290

8-14. Eastern Marketing must select a mix of advertising in order to reach a minimum of 1 million adult males, 2 million adult females, 0.5 million senior citizens, and 1.5 million children. The cost per unit and number of each type of person reached by the various advertising media are shown here. The cost per unit of advertising is: television, $200; radio, $15; magazines, $90; and newspapers, $30.

	Number of People Reached per Unit			
	Television	*Radio*	*Magazine*	*Newspaper*
Adult males	100	5	50	30
Adult females	300	20	160	5
Senior citizens	40	10	5	25
Children	100	40	10	5

Determine the number of units of each kind of advertising that will meet the standards outlined at minimum cost. What is the minimum cost?

8-15. Hardcastle Builders, Inc. builds single-family houses and condominium apartments of various sizes. The profit per unit on each of these is as follows:

Condominium *A*	$3,000
Condominium *B*	$2,800
House *C*	$3,900
House *D*	$6,200

Because of a very tight labor market, Hardcastle has not been able to increase its number of skilled employees (i.e., carpenters, bricklayers, plumbers, and roofers). The following data indicate the units of time available for each of the workers and the number of units required for each type of housing unit built.

	Units of Time Required per Unit for Each Housing Type				
Labor Type	*Condo A*	*Condo B*	*House C*	*House D*	*Units of Time Available*
Carpenters	95	110	105	160	5,000
Bricklayers	40	50	45	70	4,000
Plumbers	20	50	60	90	2,500
Roofers	25	14	30	50	1,500

a. To maximize profit, how many units of each type of housing should be built? What is the maximum profit?
b. What is the shadow price per unit for each type of labor? Assuming that the wage rate is the same for all types of labor and that the firm could add one unit of labor, what type should be hired?
c. What if the profit on house *D* falls to $2,000 per unit? How many units of each housing type should be built and what is the new maximum profit?

8-16. Mid-South Securities invests funds for a variety of institutional accounts. A new account, the Southern Teamsters Union, has $14,250,000 in cash to be invested. The union has specified that the following conditions must be met:

1. No more than 50 percent of the assets can be invested in common stock.
2. No more than 15 percent of the assets can be invested in any one common stock.
3. No more than 35 percent of the assets may be invested in Treasury bonds.

The approved securities and the rate of return on each are shown here:

Common Stock		***Fixed-Income Securities***	
High-technology stocks		U.S. Treasury bonds	9%
Ectotronics	12.5%	U.S. Treasury bills	7%
Digital Products	11.4%	United Motor bonds	10.5%
Floppy Disk Inc.	13.2%		
Other stocks			
Western Foods	10.8%		
Southern Steel	8.9%		
International Publ.	12.3%		

a. Determine the amount to be invested in a set of securities that will maximize the rate of return on the pension fund's assets while meeting the requirements just outlined. (Hint: The total amount invested cannot exceed $14,250,000.)
b. What are the dollar and percentage returns on the total investment?

8-17. The Springfield school district consists of six neighborhoods and four schools. The capacity of each school, number of students in each neighborhood, and the costs of busing one student between each neighborhood and school are shown here. If the objective of the school district is to minimize transportation cost, how should the children be assigned to the schools? How many decision variables are there, and how many can be expected to be nonzero?

	Cost per Student in Neighborhood:						
School	***1***	***2***	***3***	***4***	***5***	***6***	***School Capacity***
Adams	0.25	0.50	0.40	0.60	0.20	0.15	500
Hillcrest	0.30	0.40	0.50	0.20	0.40	0.30	1,000
Lincoln	0.40	0.30	0.20	0.60	0.35	0.40	800
Central	0.20	0.30	0.40	0.50	0.45	0.35	1,100
Total students in neighborhood	300	700	500	400	800	700	

Integrating Case Study III

Bond Construction Company

Upon completing a bachelor's degree program in business in 1992, Allen Bond opened his own construction business that specializes in building garages for residential homes. By 1998, the firm had grown substantially and employed 20 carpenters who were paid $25,000 per year. This was the entire employment complement of the firm; all managerial, accounting, and clerical functions were provided by Mr. Bond. During 1998, the company built 400 garages. Excluding materials, which are provided by the customer, each garage sells for $1,600.

The firm has a large stock of capital equipment including trucks, tools, and surveying instruments. Bond has developed a measure for a unit of capital that includes one truck and a specified amount of other equipment, including a ladder, several power tools, and an air compressor. Currently, the price to rent one of these units is $5,000 per year, and 15 units are rented. Costs of capital and labor are the only significant explicit expenses that the firm has. The labor and capital inputs can be varied daily.

Bond used a $100,000 inheritance to start the business and is quite pleased that he has received several offers in the past month to sell the firm for $300,000. The firm's accountant has prepared an income statement for 1998 that shows a profit of $40,000 after paying Bond a salary of $25,000 for the year. The market interest rate is 14 percent per year for loans to risky businesses such as this one.

Bond has just completed a managerial economics course in the evening school program of the local community college. Although he is not sure he understood everything, the class did make him aware of many problems he had not considered before. For example, is Bond Construction really earning a profit? Is the current mix of capital and labor optimal?

The pressures of managing this business are beginning to bother Bond, and he is wondering if he would not be happier simply taking a job at Hectel, Inc., a very large construction firm in the same area. Every year the personnel manager calls him with a job offer. In January 1998, the salary offer was $40,000 per year.

Bond decides to hire a consultant to make a thorough economic analysis of the firm's operation. Unfortunately, he has kept very few records that might be used for economic analysis, although records were maintained on output and inputs of capital and labor for each of the 7 years of operation. The relevant data are as follows:

Year	Output (Number of Garages)	Capital (Units)	Labor Input (Worker-Years)
1992	35	1	2
1993	49	1	3
1994	81	4	4
1995	156	4	9
1996	255	8	14
1997	277	12	14
1998	400	15	20

Bond thinks that the Cobb-Douglas function,

$$Q = AK^aL^b$$

describes the production process.

Assume that the firm has retained you as a consultant and that the information provided above is all that is available.

1. Determine the true economic profit earned by the firm in 1998.
2. If the economic profit earned by the firm in 1998 was negative, determine the breakeven output rate for the firm. Use the information developed to determine all the relevant total and per-unit cost functions.
3. Use the ordinary least-squares method to estimate the production function. Are returns to scale increasing, decreasing, or constant? Explain.
4. Determine the optimal mix of labor and capital that should be used to produce 400 units of output. Compare this mix to the actual combination of labor and capital used in 1998. *NOTE:* For a Cobb-Douglas production function, the equations for the marginal products of labor and capital are

$$MP_L = bAK^aL^{b-1}$$

and

$$MP_K = aAK^{a-1}L^b$$

PART IV
Market Structure

CHAPTER 9
Perfect Competition and Monopoly

CHAPTER 10
Monopolistic Competition, Oligopoly, and Barriers to Entry

CHAPTER 11
Game Theory and Strategic Behavior

CHAPTER 9

Perfect Competition and Monopoly

- **Preview**
- **Market Structure**
  - Number and Size Distribution of Sellers
  - Number and Size Distribution of Buyers
  - Product Differentiation
  - Conditions of Entry and Exit
- **Perfect Competition**
  - Characteristics
  - The Equilibrium Price
  - Profit-Maximizing Output in the Short Run
  - Losses and the Shutdown Decision
  - Profit-Maximizing Output in the Long Run
  - Evaluation of Perfect Competition
- **Monopoly**
  - Characteristics
  - Profit-Maximizing Price and Output in the Short Run
  - Profit-Maximizing Price and Output in the Long Run
  - Allocative Inefficiency and Income Redistribution
  - Technical Inefficiency and Rent Seeking
- **Relevance of Perfect Competition and Monopoly**
- **Summary**
- **Discussion Questions**
- **Problems**

PREVIEW

One of the most important decisions made by managers is setting the price of the firm's product. If price is set too high, the firm will be unable to compete with other suppliers. But if the price is too low, the firm may not be able to earn a normal rate of profit.

Pricing decisions are affected by the economic environment in which the firm operates. An important dimension of this environment is the degree of competition faced by the firm. A firm in a very competitive market may have little or no control over price. In that case, managerial attention must be focused on the rate of output to be produced. Conversely, a firm that is the only seller of a product may have considerable freedom in setting price. In this chapter, two market environments are considered: perfect competition and monopoly. These two cases can be thought of as extremes of market structure. For each case, economic theory is used to analyze pricing and output decisions of managers.

The first section of this chapter suggests criteria for categorizing market structures. The second and third sections discuss pricing and output decisions in perfectly competitive and monopoly market structures. The final section is a brief evaluation of the relevance of the perfect competition and monopoly models. Other market structures are considered in chapter 10.

MARKET STRUCTURE

Managers must tailor their decisions to the specific market environment in which their firms operate. For example, a manager of a business that is the patent holder and the only supplier of a new wonder drug will act differently than a manager of a firm trying to survive in the very competitive fast-food industry.

Because the decision-making environment depends on the structure of the market, it follows that no single theory of the firm can adequately describe all of the conditions in which firms operate. However, it does not follow that there must be a unique theory corresponding to every conceivable market structure. By categorizing markets in terms of their basic characteristics, it may be possible to identify a limited number of market structures that can be used to analyze decision making. Although there are many possible ways of categorizing market structures, four main characteristics are frequently employed.

Number and Size Distribution of Sellers

The ability of an individual firm to affect the price and total amount of a product supplied to a market is related to the number of firms providing that product. If there are numerous sellers of nearly equal size, the influence of any one firm is likely to be small. In contrast, in a market consisting of only a few sellers, an individual firm can have considerable impact on price and total supply.

The size distribution of firms is also an important characteristic of market structure. When the market includes a dominant firm or a few large firms that provide a substantial proportion of total supply, those large businesses may be able to exert considerable influence over price and product attributes. For example, in the market for computer software, Microsoft is the dominant firm. The products of many smaller software suppliers are designed to be compatible with those of Microsoft, and their prices are influenced by prices set by Microsoft. Conversely, in a market with firms of nearly equal size, individual sellers are likely to have less influence.

Number and Size Distribution of Buyers

Markets can also be characterized by the number and size distribution of buyers. Where there are many small purchasers of a product, all buyers are likely to pay about the same price. However, if there is only one purchaser, that buyer is in a position to demand lower prices from sellers. Similarly, if a market consists of many small buyers and one or a few firms making volume purchases, the larger firms may be able to buy at lower prices. For example, because of their sales volume, IBM and AT&T may be able to obtain electronic components at prices below those of competitors.

The market structures discussed in this chapter and in chapter 10 assume that the market has a large number of small buyers. Situations where buyers can influence price are referred to as monopsonies or oligopsonies and are discussed in chapter 13.

Product Differentiation

Product differentiation refers to the degree that the output of one firm differs from that of other firms in a market. Where products are undifferentiated, decisions to buy are made strictly on the basis of price. In these markets, sellers who attempt to charge a higher price are unable to sell their output. If there is no difference in price, the buyer has no preference as to sellers. Wheat is a good example of an undifferentiated product. Although there are several grades, all wheat of a given grade sells for the same price in a given market. Buyers are usually not told who produced the wheat, nor do they care. If properly graded, wheat from one supplier is as good as wheat from another.

At the other extreme, consider a product that is viewed by buyers as having unique characteristics. A new Rolls-Royce automobile is an example. Even the most naive car buyer would be unlikely to mistake a Rolls-Royce for a Ford. A Rolls-Royce has come to represent the ultimate in automobile luxury. As such, it commands a price that may be 10 to 15 times that of a new Ford.

Product differentiation is an important market characteristic because it indicates a firm's ability to affect price. If a firm's product is perceived as having unique features, it can command a premium price. However, products considered less desirable will be purchased only if the seller is willing to accept a lower price. For example, consumers will pay extra for fresh San Francisco sourdough bread, but will buy day-old Wonder Bread only if the price is substantially reduced.

Conditions of Entry and Exit

Ease of entry and exit are crucial determinants of the nature of a market in the long run. When it is extremely difficult for new firms to enter, existing firms will have much greater freedom in making pricing and output decisions than if they must be concerned about new entrants who have been attracted by the lure of high profits. Consider a drug manufacturer that holds a patent that prohibits other firms from making the drug. If there are no close substitutes for the product, that firm will essentially be free from competition now and for the duration of the patent. Thus, its managers can make pricing decisions without worrying about losing market share to new entrants. However, if the drug can be easily copied, and if prices are substantially above costs, new firms may enter the market.

Ease of exit also affects managerial behavior. Suppose that certain firms in a market have been earning less than the normal rate of profit. If the resources used to produce the product can easily be transferred from one use to another, some of those resources will

be shifted to other industries, where they can earn a higher rate of return. However, if the resources are highly specialized, they may have little value in another industry. For example, the track and terminals of an unprofitable railroad may have few alternative uses, and may only be sold for their salvage value. This makes exit more difficult and costly.

Key Concepts

- In markets where there are a large number of small buyers and sellers, individual firms have little control over price.
- By differentiating its product, a firm can gain some control over price.
- If it is easy for new firms to enter an industry, existing firms may have little freedom in their pricing decisions.

PERFECT COMPETITION

The term *perfect competition* is something of a misnomer. In a perfectly competitive world there really is no overt competition between economic units. As buyers and sellers make business decisions, they do not have to take into account the effect of their actions on other participants in the market. The reason is that the individual economic units in perfect competition are so small relative to the total market that their actions have no perceptible impact on other buyers and sellers. Hence, decisions can be made without considering the reactions of others. In perfect competition, market participants do not compete against one another. Rather, they make decisions in an economic environment that they perceive as being fixed or given.

Characteristics

The concept of perfect competition can be defined in terms of the market structure characteristics of the preceding section. First, there must be a large number of sellers in the market with no single seller able to exert significant influence over price. This criterion is sometimes described in terms of sellers being *price takers* who can sell all that they can produce at the market-determined price. Graphically, this situation is depicted as sellers facing a horizontal demand curve. Similarly, the second requirement for perfect competition is that there be a large number of small buyers, each buyer being unable to influence price. That is, all buyers are price takers.

Third, perfect competition assumes easy entry and exit from an industry. If price is above cost, resulting in economic profits, resources can be mobilized to create new firms or to expand the production capacity of firms already in the industry. If profits are below average, resources can easily be transferred from the industry and used to produce other products at higher profit rates. Finally, under perfect competition, it is assumed that the product is totally undifferentiated. One firm's output cannot be distinguished from that of other producers. As a result, purchasing decisions are based entirely on price. If the firm sets its price above the market-determined level, it will be unable to attract buyers. Price cutting is unnecessary because producers can sell their total output at the market price. Characteristics of perfectly competitive markets are summarized in Table 9.1.[1]

[1]Sometimes the assumption of perfect knowledge regarding prices and technology is included as a fifth characteristic of perfect competition. For ease of exposition, it is not included here.

Number and size distribution of sellers	Many small sellers. No seller is able to exert a significant influence over price.
Number and size distribution of buyers	Many small buyers. No buyer is able to exert a significant influence over price.
Product differentiation	Product undifferentiated. Decisions to buy are made on the basis of price.
Conditions of entry and exit	Easy entry and exit. Resources are easily transferable among industries.

The Equilibrium Price

In the preceding section, reference was made to the market-determined price. Although no single entity in a perfectly competitive market can affect price, the aggregate effect of the participants in the market is important in price determination. Indeed, the interaction of supply and demand determines the equilibrium price and the quantity to be exchanged.

Consider a hypothetical market for wheat. Each wheat producer has an individual supply schedule. Two such schedules are shown in Table 9.2. These schedules indicate the quantity of wheat that will be produced per period at different wheat prices. At each price the decision rule is the same: Additional wheat will be supplied only if the price is high enough to allow the supplier to earn at least a normal rate of profit on the incremental output. Table 9.2 shows that higher anticipated prices are necessary to induce the producers to supply more wheat.

For the moment, assume that the two supply schedules shown in Table 9.2 represent the only suppliers of wheat in the market. By adding the amount that each producer will provide at each price, the market supply schedule for wheat can be computed. This information appears in the fourth column of Table 9.2. For example, at a price of $6 per bushel, the first producer will supply 9,000 bushels of wheat per period and the second will supply 7,000 bushels per period. Thus, the quantity supplied to the market at $6 is 16,000 bushels of wheat per period.

Now suppose that there are 10,000 wheat producers with supply schedules as shown in column 2 and another 10,000 with schedules like that of column 3. Thus, the quantity of wheat supplied per period at each price will be 10,000 times the amounts shown in

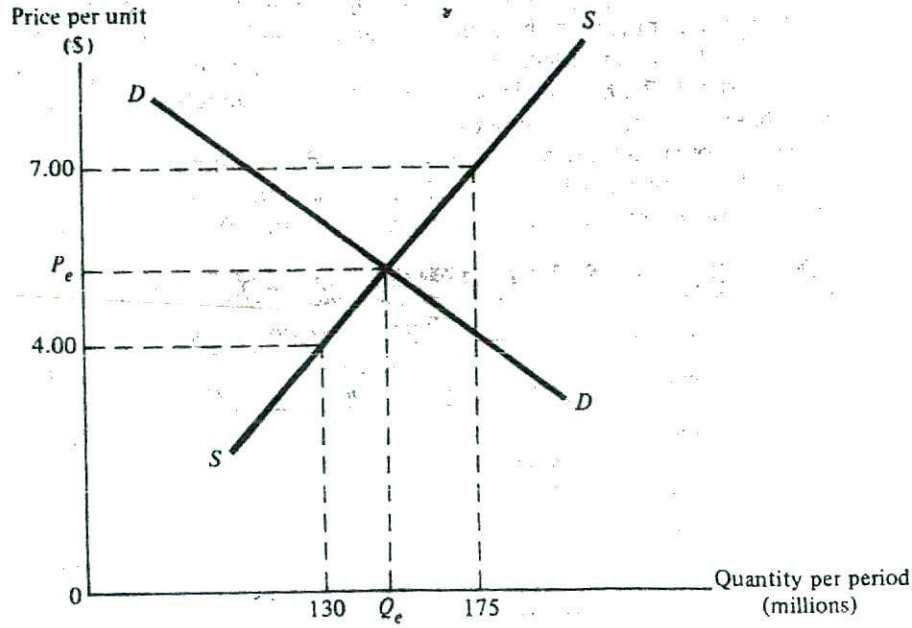
	Quantity Supplied per Period				
Price per Bushel	*Firm 1* +	*Firm 2* =	*Two-Firm Supply* ×	*10,000* =	*Total Market Supply*
$8	10,000	9,000	19,000		190 million
7	9,500	8,000	17,500		175 million
6	9,000	7,000	16,000		160 million
5	8,500	6,000	14,500		145 million
4	8,000	5,000	13,000		130 million
3	7,500	4,000	11,500		115 million

column 4 of Table 9.2, and the market supply schedule will be as shown in the last column of the table.

The supply data can be plotted to form the supply curve shown in Figure 9.1. The market demand curve is also shown. As discussed in chapter 3, the market demand is the horizontal sum of the demands of individual buyers. The equilibrium price of wheat is P_e and is determined by the point of intersection of the supply-and-demand curves. If price is greater than P_e, there is excess supply. Producers will respond by cutting prices in order to sell the excess wheat. As the price falls, quantity demanded increases and quantity supplied decreases. Alternatively, excess demand exists when the price is below P_e. This causes consumers to bid up the price of wheat. As the price increases, buyers reduce their purchases and suppliers increase production. These forces continue to operate until supply and demand come into balance at the equilibrium price, P_e.

Consider the impact of any one seller on the market supply curve. Table 9.2 shows that at a price of $6, the amount supplied will be 160 million bushels of wheat per period. Suppose that a business like firm 1 in the table decides not to produce. The effect would be to reduce supply, but only by 9,000 bushels or about six-thousandths of 1 percent of total supply. Thus, the action of a single seller would have no measurable impact on the market. Although the supply function would shift to the left, the shift would be nearly impossible to detect and would have virtually no impact on the equilibrium price, P_e.

Now consider the effect of individual consumer demands. Suppose that there are 20,000 buyers, each with the same demand schedule. Should one buyer drop out of the market, the demand curve would shift to the left. But the shift would be so small as to have no observable impact. Once again, the market price would be essentially unaf-



fected. Thus, it is seen that in markets with large numbers of buyers and sellers, individual firms and consumers are unable to affect price. That is, they are price takers in the market.

Case Study
Credit Cards and Perfect Competition

In 1997, over $700 billion purchases were charged on credit cards, and this total is increasing at a rate of over 10 percent per year. At first glance, the credit card market would seem to be a rather concentrated industry. Visa, MasterCard, and American Express are the most familiar names, and over 60 percent of all charges are made using one of these three cards. But, on closer examination, the industry seems to exhibit most of the characteristics of perfect competition.

Consider first the number and size distribution of buyers and sellers. Although Visa, MasterCard, and American Express are the choice of the majority of consumers, these cards do not originate from just three firms. In fact, there are over 6,000 enterprises (primarily banks and credit unions) in the United States that offer charge cards to over 90 million credit card holders. One person's Visa card may have been issued by his company's credit union in Los Angeles, while a next-door neighbor may have acquired hers from a Miami bank when she was living in Florida.

Credit cards are a relatively homogeneous product. Most Visa cards are similar in appearance, and they can all be used for the same purposes. When a charge is made, the merchant is unlikely to notice who it was that actually issued the card. Entry into and exit from the credit card market is easy, as evidenced by the 6,000 institutions that currently offer cards. Although a new firm might find it difficult to enter the market, a financially sound bank, even one of modest size, could obtain the right to offer a MasterCard or Visa card from the parent companies with little difficulty. If the bank wanted to leave the field, there would be a ready market to sell its accounts to other credit card suppliers.

Thus, it would seem that the credit card industry meets most of the characteristics for a perfectly competitive market. However, in some ways the industry appears not to behave in a manner consistent with the model of perfect competition, which is developed in the following sections. This anomaly will be explained in a later case study in this chapter. ■

Profit-Maximizing Output in the Short Run

This section analyzes the profit-maximizing output of a profitable competitive firm in the short run. As discussed in chapter 6, the short run is defined as a period of time in which at least one input is fixed. Often, the firm's capital stock is viewed as the fixed input.

Accordingly, this analysis assumes that the number of production facilities in the industry and the size of each facility do not change, because the period being considered is too short to allow businesses to enter or leave the industry or to alter the basic nature of their operations. The period of time that can properly be designated as the short run depends on the characteristics of the industry. For production of electric power, it may take as much as ten years to bring a new generating plant on line. In contrast, economic profits in service industries may attract new entrants in a matter of weeks.

Demand The firm in perfect competition faces a horizontal demand curve at the market price for its product. This can be seen by evaluating the effect of the firm's decisions on market demand. Using wheat as an example, let the market demand equation be given by

$$Q_D = 170{,}000{,}000 - 10{,}000{,}000P \tag{9-1}$$

The equation implies that quantity demanded per period is reduced by 10 million bushels per dollar increase in price. Suppose that the supply equation is given by

$$Q_s = 70{,}000{,}000 + 15{,}000{,}000P \tag{9-2}$$

Equation (9-2) corresponds to the market supply data from Table 9.2 and indicates that a \$1 price increase results in 15 million extra bushels of wheat being supplied per period.

Equating these supply-and-demand functions and solving for P and Q yield the equilibrium values. Specifically, price equals \$4 and quantity is 130,000,000 bushels per period. Now assume that a supplier like firm 2, as shown in Table 9.2, leaves the market. The table implies that the supply equation for that single supplier is given by the equation $q_2 = 1{,}000 + 1{,}000P$. Subtracting q_2 from equation (9-2) gives a new equation for market supply:

$$Q_s' = 69{,}999{,}000 + 14{,}999{,}000P \tag{9-3}$$

Solving equations (9-1) and (9-3) gives $P = \$4.0002$ and $Q = 129{,}998{,}000$. Note that the exit of one producer increased the equilibrium price by \$0.0002 and reduced quantity by 2,000 bushels.

Thus, it is seen that the output decisions of individual producers have no significant impact on the market price. If the one supplier remains in the market, the equilibrium price will be \$4. But if the small producer leaves the market, the price is still very close to \$4. Graphically, this is portrayed by a horizontal demand curve at the \$4 equilibrium price, as shown in Figure 9.2. The curve indicates that an individual firm can sell as much as it can produce at the given price. However, if the firm sets its price greater than at \$4, it will have no sales because consumers will purchase from other suppliers. Conversely, there is no reason to sell below \$4 because the firm's total output can be sold at the market price of \$4 per bushel.

Costs It is assumed that the firm has U-shaped average and marginal cost curves, as shown in Figure 9.2. The figure shows that as quantity increases from 0 to q_m units, average cost declines and then increases beyond that point. It is important to remember that the cost curves of Figure 9.2 include a normal rate of profit. Thus, any time that the firm's price is greater than average cost, it is earning economic profit.

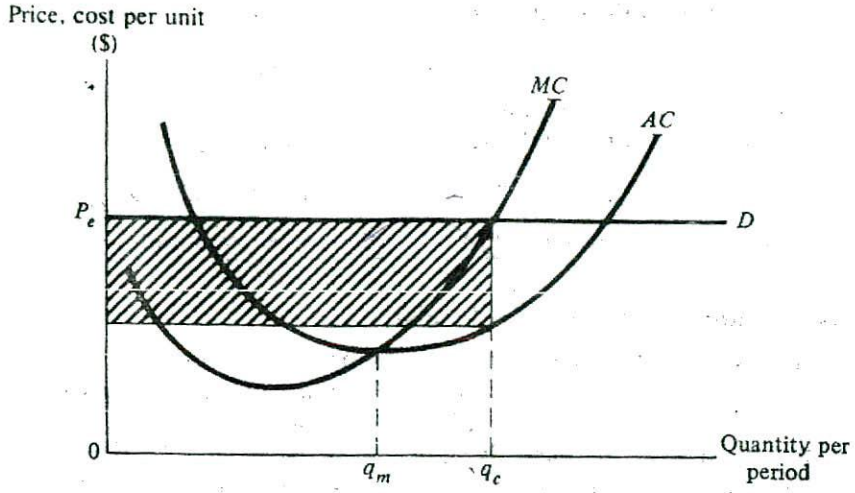


FIGURE 9.2 Short-Run Profit-Maximizing Output in Perfect Competition

Equilibrium Output Because price is determined in the market and the product is homogeneous, the only decision left to the manager of a firm in a perfectly competitive market is how much output to produce. The profit-maximizing output is determined where the extra revenue generated by selling the last unit (i.e., the market price) just equals the marginal cost of producing that unit. For a horizontal demand curve such as that of Figure 9.2, this condition is met by increasing the rate of production to q_c where price equals marginal cost. If the firm increases output beyond this point, the additional revenue, P_e, is less than the extra costs as shown by the marginal cost curve. In contrast, if production is reduced below q_c, the loss of revenues is greater than the reduction in costs, and profits decrease. The output rate, q_c, represents the short-run equilibrium for the competitive firm in the sense that a profit-maximizing manager has no incentive to alter output as long as the demand and cost curves remain unchanged.

Example Pizza Profits

A new pizza place, Fredrico's, opens in New York City. The average price of a medium pizza in New York is $10 and, because of the large number of pizza sellers, this price will not be affected by the new entrant in the market. The owner of Fredrico's estimates that monthly total costs, including a normal profit, will be

$$TC = 1{,}000 + 2Q + 0.01Q^2$$

To maximize total profit, how many pizzas should be produced each month? In the short run, how much economic profit will the business earn each month?

Solution Taking the derivative of the total cost equation with respect to Q gives the marginal cost equation

$$MC = \frac{dTC}{dQ} = 2 + 0.02Q$$

Profit is maximized by equating price and marginal cost. Thus, the profit-maximizing output is given by the solution to

$$10 = 2 + 0.02Q$$

which is 400 pizzas per month.

Economic profit is total revenue minus total cost, or

$$TR - TC = 10(400) - [1{,}000 + 2(400) + 0.01(400^2)] = \$600$$

Thus, in the short run, economic profit will be $600 per month.

Losses and the Shutdown Decision

Simply because profit maximization is the objective of managers is no guarantee that a firm will actually earn economic or even normal profits. Oversupply, poor management, or high costs may prevent a firm from operating profitably at any rate of output. That is, maximum profit may actually be negative.

The course of action adopted by managers of an unprofitable firm should be based on a consideration of the alternatives. One option would be to continue producing at the least unprofitable (i.e., smallest loss) rate of output. Another would be to shut down operations and produce nothing. The best choice is the alternative that minimizes the firm's losses.

In the short run, the consequences of shutting down versus continuing production are illustrated using the hypothetical data for a firm in perfect competition found in Table 9.3. Column (1) shows various rates of output that could be produced, and column (2) shows total fixed costs. By definition, these fixed costs are constant for all rates of output. In column (3) total variable costs are reported. The remaining data in the table have been computed from the information in the first three columns.

Recall that total costs are the sum of total variable costs and the total fixed cost. Marginal cost is the change in total cost (or total variable cost) resulting from a one-unit change in output. Average variable cost is total variable cost divided by quantity.

TABLE 9.3 Short-Run Output and Cost Data

Quantity	*Total Fixed Cost*	*Total Variable Cost*	*Total Cost*	*Marginal Cost*	*Average Variable Cost*	*Average Total Cost*
0	$5	$ 0	$ 5	—	—	—
1	5	5	10	$5	$5.00	$10.00
2	5	9	14	4	4.50	7.00
3	5	12	17	3	4.00	5.67
4	5	14	19	2	3.50	4.75
5	5	17	22	3	3.40	4.40
6	5	21	26	4	3.50	4.33
7	5	26	31	5	3.72	4.42
8	5	32	37	6	4.00	4.63
9	5	39	44	7	4.33	4.88

Similarly, average total cost is total cost divided by quantity. Note that marginal, average variable, and average total costs first decrease and then increase. This is consistent with the discussion of costs in chapter 7.

Now consider the optimal rate of output for a profit-maximizing firm facing a horizontal demand curve. The optimal quantity depends on the market-determined price. The decision rule is that the firm should produce an additional unit of output if the selling price is at least as great as the marginal cost of production.

For example, if the price is \$5, the firm should produce seven units because the marginal cost of the seventh unit is \$5. If price increases to \$6, the optimal quantity would be eight units because the marginal cost of the eighth unit is \$6. Similarly, at a price of \$7, nine units should be produced.

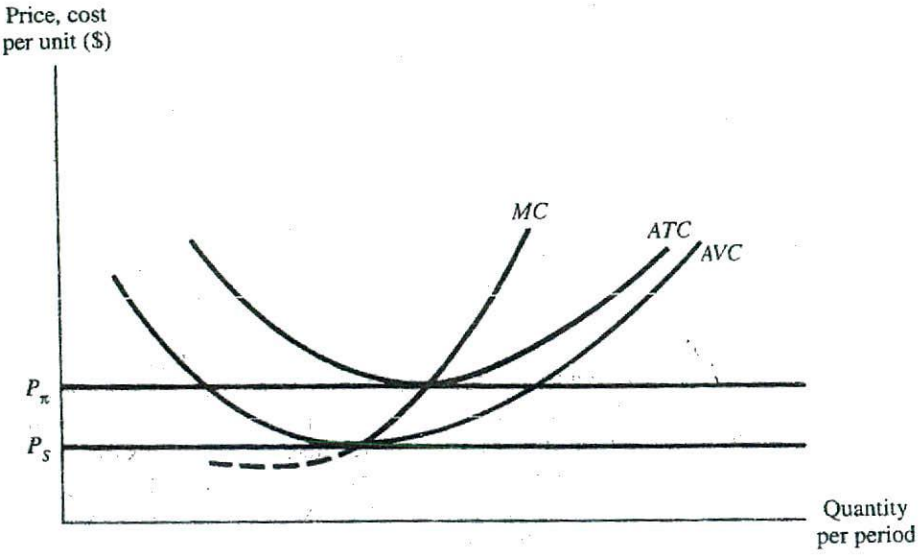
What if the price declines to \$3? Applying the same logic, it would seem that the firm should produce five units. But note that average variable cost at five units is \$3.40 and total variable cost is \$17. This \$17 is an expense that could be avoided if the firm did not produce the five units of output. Because the firm sells output for \$3 per unit, its total revenue is \$15. Hence, by producing, the firm adds \$15 to its total revenue but incurs additional (and avoidable) costs of \$17. Adding the fixed cost of \$5 to the avoidable loss of \$2 results in a total loss of \$7.

At a price of \$3, producing at any other output rate would cause equal or greater losses. For example, cutting back to four units would also result in a total loss of \$7, and expanding output to six units would increase the firm's loss to \$8. However, the firm's managers do have one other option. They could shut down the firm's operations and produce nothing. In this case, there would be no revenue and no variable costs. The loss would be the \$5 in fixed costs that must be paid whether or not the firm produces. Thus, by shutting down, the firm loses \$5, compared to a minimum loss of \$7 if any production takes place. In general, the decision rule is that a firm minimizes its losses by shutting down when price drops below average variable cost.

Now, suppose that the firm can sell at a price of \$4. Price equals marginal cost at a quantity of six units. Sale of six units will generate revenues of \$24 and cause the firm to incur total costs of \$26, for a net loss of \$2. At any other rate of output, the losses are even greater. Is the optimal choice again to shut down, as it was at a price of \$3?

If the firm shuts down, it must still pay the fixed costs of \$5. However, by producing six units, the loss is only \$2. Clearly, the firm minimizes its losses by continuing to produce. The key to the decision is an examination of price in relation to average variable cost. As long as price exceeds average variable costs, the firm is better off if it continues to produce. The reason is that revenue will be sufficient to cover variable costs and make a contribution to payment of the firm's fixed costs. In contrast, shutting down means that the firm's loss is the entire fixed cost.

This concept is illustrated in Figure 9.3. The portion of the marginal cost curve that lies above average variable cost represents the firm's supply curve. That is, it shows the profit-maximizing output at each price. When price drops below average variable cost (i.e., below P_3), the firm minimizes its losses by shutting down. If price is greater than average variable cost but less than average total cost, the firm earns less than a normal rate of profit but loses less than if its operations were shut down. Finally, for prices greater than or equal to average total cost (i.e., P_π or greater), the firm earns at least a normal rate of profit.



Case Study

Texas Instruments' Exit from the Home Computer Market

The unfortunate experience of Texas Instruments (TI) in the home computer industry during the early 1980s illustrates the relationship between variable costs and the decision to produce or shut down. In mid-1980, TI was selling its basic home computer for $650. But cost-reducing technology and intense competition from firms such as Atari, Commodore, and Tandy drove market prices steadily downward. By mid-1982, TI had been forced to cut its price to $249, and by early 1983, the firm's basic computer could be purchased for just $149. Still, because TI's variable cost per unit was about $100, the firm was better off continuing to produce, even at this low price.

But improved technology and competitive pressures continued to push prices even lower. By the fall of 1983, market conditions required TI to reduce its price to $99. Because this level was less than the $100 variable cost, the firm stopped producing home computers. However, at the time of the shutdown decision, the company had almost 500,000 unsold computers on hand. Texas Instruments finally got out of the home computer business by dumping its remaining inventory onto the market at $49 per unit.

If the variable cost of producing a computer was $100, why would the firm be willing to sell its remaining machines for less than that amount? The explanation is that the $100 was a variable cost before the computers were produced, but a sunk cost afterward. Thus, the firm was better off selling the computers than keeping them as long as the $49 price was greater than the transportation and marketing expenses (the costs that were still variable) of selling each unit.

Texas Instruments was not alone in its exit from the home computer market. Dozens of smaller companies also determined that their losses would be less if they shut down than if they continued to produce. ■

Two qualifications apply to the basic decision rule. First, the rule does not necessarily mean that managers should shut down operations every time price drops below average variable cost. In many cases, substantial costs are incurred when a production process is shut down and also when it is restarted. For example, in steel manufacturing, several days may be required to bring a blast furnace up to operating temperature, and there are costs involved in laying off and recalling workers. Also, a firm that shuts down and then reopens may find that its customers are buying from other suppliers. These costs must be taken into account. They suggest that a decision to shut down will be made only if it is expected that price will remain below average variable cost for an extended period of time.

The second qualification involves the distinction between the short run and the long run. Note that the decision to shut down depends on whether the firm can make a contribution to its fixed cost by continuing to produce. But in the long run, there are no fixed costs. Buildings can be sold, equipment can be auctioned off, and purchase contracts will expire. Thus, in the long run, if price is expected to remain below average total cost, the firm will shut down and go out of business. Basically, the same decision rule applies to both the short and the long run—a firm should continue to produce as long as revenues exceed variable or avoidable costs.

Example **Calculating the Shutdown Price**

A bicycle manufacturer faces a horizontal demand curve. The firm's total costs are given by the equation

$$TVC = 150Q - 20Q^2 + Q^3$$

where Q is quantity.

Below what price should the firm shut down operations?

Solution Marginal cost is the derivative of total cost with respect to quantity. Thus

$$MC = \frac{dTVC}{dQ} = 150 - 40Q + 3Q^2$$

The average variable cost equation is given by

$$AVC = \frac{TVC}{Q} = \frac{150Q - 20Q^2 + Q^3}{Q} = 150 - 20Q + Q^2$$

The shutdown point is where price equals minimum average variable cost. But profit maximization requires that price also equal marginal cost. Thus, by setting $MC = AVC$, the result is

$$150 - 40Q + 3Q^2 = 150 - 20Q + Q^2$$

Rearranging terms gives

$$2Q^2 - 20Q = 0$$

which can be rewritten as

$$2Q(Q - 10) = 0$$

Solving this equation gives $Q = 0$ or $Q = 10$. Substituting $Q = 10$ into the marginal cost equation gives

$$P = MC = 150 - 40(10) + 3(100) = 50$$

A similar substitution for $Q = 0$ yields $P = 150$. The relevant solution is the nonzero output. Thus, if the price falls below \$50 per unit, the firm should shut down.

Key Concepts

- The firm in perfect competition maximizes profit by producing at the rate of output where price equals marginal cost.
- In the short run, managers of a firm should shut down the operation if price is below average variable cost.
- If price is greater than average variable cost but less than average total cost, the firm should continue to produce in the short run because a contribution can be made to fixed costs.

Profit-Maximizing Output in the Long Run

A key characteristic of the perfect competition model is ease of entry and exit. However, this assumption does not imply that such changes are instantaneous. It takes time for new firms to build facilities and for existing firms to increase output. Similarly, firms leaving an industry may experience delays in converting their resources to other uses. These problems of entry and exit are not considered in the short-run analysis.

In the long run, all inputs are variable. Firms can enter or exit an industry and can also change the size of their production facilities. As a result, although the output rate q_c in Figure 9.2 on page 315 represents the profit-maximizing decision in the short run, it may not be the optimal choice in the long run. Producing at q_c, the firm is earning economic profit. In the figure, per-unit economic profits are given by the vertical distance between the average cost curve and the demand curve at the output rate q_c. Total economic profit is shown by the shaded area. Because the average cost curve already includes a normal profit rate, the implication is that capital invested in the firm is earning substantially more than capital used in other sectors of the economy. Thus, owners of capital have an incentive to withdraw their capital from those sectors yielding only a normal return and to employ it in this industry where greater profits can be earned.

As additional capital flows into the industry, more output will be produced at each price. Thus the market supply curve, SS, shifts to the right to $S'S'$, as shown in Figure 9.4. This shift may result from more firms operating in the industry or the facilities of existing firms being expanded. It is useful to think of the supply shift as indicating that more of the product will be produced at any given price than before the inflow of capital. As the supply curve shifts to the right, the intersection of supply and demand causes a new

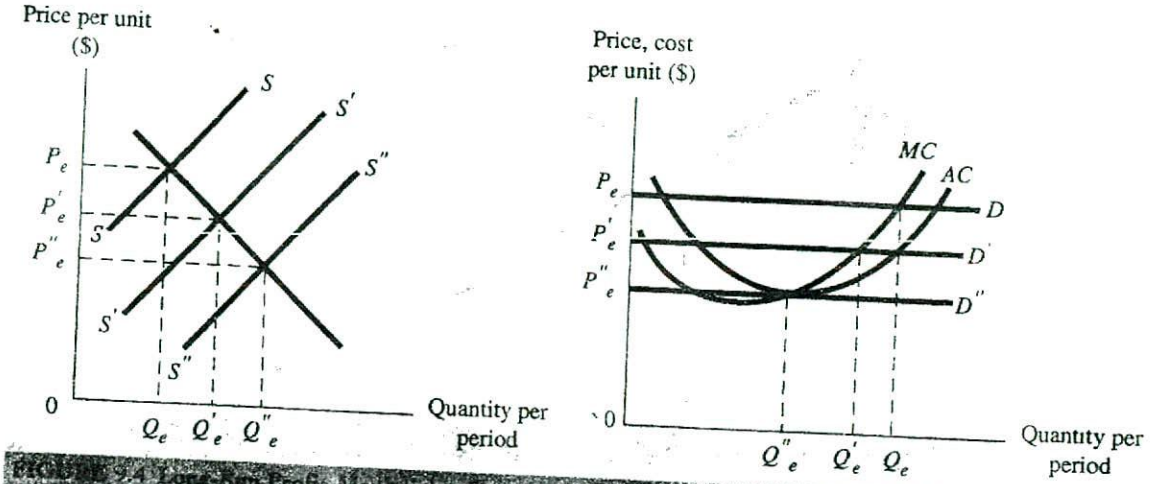


FIGURE 9.4 Long-Run Profit-Maximizing Output

equilibrium price, P'_e. This result is shown in Figure 9.4. At the lower price, the individual firm now faces a new horizontal demand curve, D'. But at the price, P'_e, the output rate Q_c no longer maximizes profit. At Q_c marginal cost is greater than incremental revenue. Now the firm maximizes profits by reducing the rate of output to Q'_c, where price again is equal to marginal cost.

Producing Q'_c units per period and selling at P'_e, the firm is less profitable than before, but it is still earning economic profit. This can be seen by observing that the firm's average revenue, P'_e, is greater than average cost at the output rate Q'_c. Thus, there is an incentive for additional capital to flow into the industry. This additional capital expands capacity and causes further rightward shifts of the industry supply curve. The inflow of capital will continue until the supply curve is shifted to $S''S''$ and the equilibrium price is reduced to P''_e. Hence, the demand curve faced by the individual firm is shown by curve P''_eD'' in Figure 9.4. In this situation, profit is maximized by producing Q''_c. Notice that at Q''_c, price is equal to marginal cost, but price is also equal to average cost. Thus the firm's average revenue just equals average cost. Hence, the firm is earning a normal rate of profit, but there is no economic profit.

Because the return to capital in the industry is no higher than the return earned in other segments of the economy, there is no further incentive for capital to flow into the industry. However, because capital earns at least a normal return, there is no reason for owners to withdraw capital from the industry. Hence the output rate Q''_c, where price equals average cost, is the long-run equilibrium for the representative firm in this perfectly competitive industry.

Evaluation of Perfect Competition

Prices play a central role in economic theory. The price that a person is willing to pay for a good or service is a measure of the value attached to having one more unit of that product. If a person is willing to buy at a specified price, the implication is that

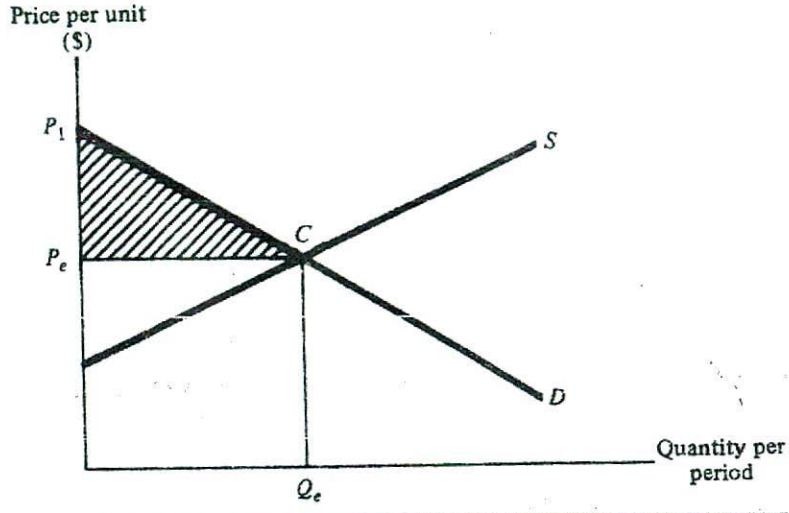


FIGURE 9.5 Consumer Surplus

nothing else could provide equal satisfaction for the same amount of money. Obviously, individuals differ in their valuation of products. Also, preferences are affected by how much of a good or service the person already has. Usually, additional units are considered less valuable than the initial units purchased.

Preferences for goods and services are depicted by a demand curve such as in Figure 9.5. The curve indicates the maximum amount that anyone will pay for each additional unit. For example, the curve shows that there is a consumer who would pay P_1 for the first unit, someone who would pay almost that price for the next unit, and so on.

The market equilibrium price, P_e, in Figure 9.5 is determined by the intersection of supply and demand. But only those consumers who value the product at least P_e (i.e., that part of the demand curve above P_e) will buy it at that price. Thus quantity demanded will be Q_e. But in perfect competition, everyone pays the same price. The implication is that all those who valued the product more than P_e receive more benefit than they paid for. This extra benefit is referred to as the consumer surplus. For each unit purchased, the consumer surplus in Figure 9.5 is the vertical distance between the demand curve and P_e. The total consumer surplus is the area of the shaded triangle, P_1CP_e.

Marginal costs are a measure of the opportunity cost of producing one more unit of the product. For example, to produce an additional automobile, fuel, labor, and capital must be diverted from other uses. The value of these inputs in those other uses is measured by their cost. The sum of these input costs is the marginal cost and represents the opportunity cost of producing an additional car, as shown by the supply curve in Figure 9.5.

The profit-maximizing firm in perfect competition will expand production until price equals marginal cost. Concurrently, buyers will purchase the firm's product until price exceeds the relative value that they attach to the product. Because the price paid

by the consumer is identical to the additional revenue received by sellers, the equilibrium output in perfect competition has the following characteristic: The value that the last buyer attaches to the last unit of output produced is just equal to the opportunity cost of producing that unit of output.

It is often suggested that perfect competition results in the right amount of the product being produced. The preceding argument is the justification for that statement. If the value of the last unit of output is less than the cost of its production, social welfare would be improved by shifting the resources used in producing that last unit to production of some other good or service. Conversely, if there are potential customers who attach a value to the product greater than its marginal cost and who are not being served because of insufficient production, resource allocation could be improved by increasing output.

Another way of thinking about this result is in terms of voluntary exchange. Firms and consumers will exchange goods and services only as long as both parties benefit from the trade. Thus, increased voluntary exchange implies improved resource allocation. Voluntary exchange is maximized under perfect competition because output is increased until there are no consumers willing to pay the opportunity cost of producing an additional unit of the good.

A second consequence of perfect competition is that resources are efficiently allocated among alternative uses. Consider economic profits in long-run equilibrium. In the long run, individual firms earn no more than a normal rate of profit. If the rate of profit measures the productivity of resources in a given use, then whenever resources are earning a higher rate of return in one use than in another, resource allocation could be improved by shifting resources to the higher-return use. This is exactly what occurs in the long-run model of perfect competition. Resources flow into the competitive industry (entry) until economic profit is eliminated. Conversely, if the return in the competitive sector is less than the normal rate of return, resources leave the industry (exit) until the remaining resources are earning a normal rate of return.

A third characteristic of perfect competition in the long run is that production occurs at minimum cost. Recall that the profit-maximizing output is at the minimum point on the average cost curve. This result does not mean that competitive firms necessarily are more efficient than firms in other types of market structure. However, it does imply that, given the technology available to the firm, economic forces in perfect competition require producers to minimize the per-unit cost of production.

Key Concepts

- In the long run, economic profit is eliminated by the entry of new firms. The profit-maximizing rate of output occurs where price equals both marginal and average cost.
- Consumers who would have been willing to pay more than the market price receive a consumer surplus when they buy the product.
- In perfectly competitive markets, (1) the value of the last unit exchanged equals the opportunity cost of producing it, (2) capital moves to its highest valued use, and (3) production takes place at the minimum point on the average cost curve.

Case Study
Credit Cards, Perfect Competition, and High Interest Rates

In an earlier case study in this chapter, it was argued that the credit card industry fulfills most of the criteria for a perfectly competitive market. But at least one aspect of the industry seems inconsistent with the competitive model just described.

Credit cards serve as a medium of exchange by allowing consumers to charge purchases rather than pay cash at the time of the transaction. If payment is made within 30 days, there is no finance charge. They also are a source of credit, whereby people can defer payment for a purchase for an extended period of time by paying interest. It is the high interest rates charged by credit card issuers that require explanation.

Economic theory suggests that competition should drive interest rates down as card offerers compete with one another for accounts. But the evidence of the 1980s and 1990s suggests that credit card interest rates remained high while other rates were declining. In 1998, finance charges to credit card holders were typically 13 to 14 percent, while banks were paying just 4 to 5 percent on money deposited in savings and money market accounts. Why didn't competition among the suppliers of credit cards cause finance charges to adjust to a level consistent with the cost of money to banks?

The answer involves how consumers use their charge cards. Less than one-half of credit card holders actually pay finance charges in a given month and another one-fourth do not anticipate that they will have a balance that will require that they pay interest. Consequently, only the remaining one-fourth of card holders who expect to pay finance charges are likely to base their decision of which card to select on the interest rate that must be paid. The rest of the credit card users are more concerned about other features, such as where the card is accepted, the annual fee, and the credit limit.

Thus, in setting the interest rate for credit card accounts, banks will focus on the risk associated with those who frequently use their cards as a credit instrument. But who are these people likely to be? Often, those who incur substantial finance charges on their credit cards will be consumers with a relatively high risk of default—people who cannot get credit on favorable terms elsewhere, those who have difficulty managing money, or those with relatively low net worth. Hence, the explanation for high interest rates on credit card balances is that the rates are set to account for the risk associated with the one-fourth of card holders who are most likely to borrow. The majority of customers are largely unaffected by the high interest rates. ■

MONOPOLY

Although conditions facing a monopolist are much different from those of firms in perfect competition, the two types of firms have at least one thing in common—they do not have to compete with other individual participants in the market. Sellers in perfect com-

Number and size distribution of sellers	Single seller
Number and size distribution of buyers	Unspecified
Product differentiation	No close substitutes
Conditions of entry and exit	Entry prohibited or difficult

petition are so small that they can ignore each other and consider the market environment as given. At the other extreme, the monopolist is the only seller in the market and has no competitors.

Characteristics

Monopoly can be described in terms of the market structure characteristics discussed earlier in the chapter. First, there is only one seller in the market. This means that the demand curve faced by the monopolist is the downward-sloping demand curve for the market. Second, for a firm to continue as a monopoly in the long run, there must be factors that prevent the entry of other firms. Such barriers to entry are discussed in chapter 10. Finally, the product of the monopolist must be highly differentiated from other goods. That is, there must be no good substitutes. The market structure characteristics of a monopoly are listed in Table 9.4.

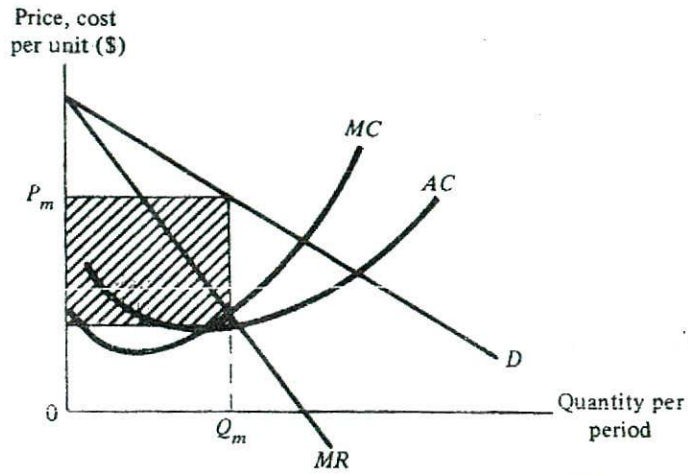
Consider a small, isolated community that has only one supplier of concrete. Essentially, that firm has a monopoly position. When residents want concrete for foundations of new houses, they will have to buy from this monopolist. The high cost of transporting concrete makes it unlikely that concrete producers in other cities will be viable competitors. At the same time, there are few good substitutes for concrete foundations. Wood, stone, and cinder block are possibilities, but they are not as strong or as easy to use as concrete.

Profit-Maximizing Price and Output in the Short Run

Demand and cost curves for a monopolist are shown in Figure 9.6. As with the perfectly competitive firm, the cost curves depict first decreasing and then increasing average costs.

Because they face a horizontal demand curve, managers of firms in a perfectly competitive world have no control over price. They simply choose the profit-maximizing output. However, because the monopolist has a downward-sloping demand curve, as shown in Figure 9.6, managers must recognize that their output decisions can influence price and vice versa. Because price must decrease in order to increase sales, an increase in output will require that the firm sell at a lower price. The effect of output changes on total revenues depends on the marginal revenue curve shown in Figure 9.6. If marginal revenue is negative, total revenue is reduced by the increased output.

The criterion for maximizing profits is the same for the monopolist as for firms in perfect competition—output should be increased until the additional revenue equals the marginal cost. For the competitive firm, the price is unaffected by output, so the decision criterion is to produce until price equals marginal cost. For the monopolist, the



equivalent criterion is to produce at Q_m in Figure 9.6, where marginal revenue equals the marginal cost. At this output, the monopolist charges what the market will allow, as indicated by the demand curve. In Figure 9.6, this price is P_m.

Profit-Maximizing Price and Output in the Long Run

Notice that producing Q_m units of output, the monopolist is earning economic profit, as indicated by the shaded area in Figure 9.6. If it were possible, other firms would enter the market to take advantage of the high rate of return. With other sellers in the market, the demand curve faced by the monopolist no longer would be the market demand curve. The firm's new demand curve would be relatively more elastic because the firm's output would represent a smaller share of total market sales and thus have a smaller effect on price. At the same time, part of the market and some of the economic profit earned by the monopolist would be captured by the new entrants. Ultimately, the market structure might evolve to an oligopoly (a small number of sellers) or even approach perfect competition. However, if the firm's monopoly position is the result of its control over scarce inputs such as mineral reserves, patents, unique managerial talent, or a choice location, entry by other firms may be impossible, and the firm will maintain its monopoly position. In this case, economic profits may persist indefinitely. Thus, Figure 9.6 may depict both the short-run and the long-run profit-maximizing price and output for a monopoly.

Key Concepts

- As the only seller, a monopolist faces the market demand curve. The profit-maximizing output is determined by the point where marginal revenue equals marginal cost.
- If entry by other firms is difficult, even in the long run, the monopolist can earn economic profits.

Example **Computing Profit-Maximizing Price and Output for a Monopolist**

Suppose that the total cost equation (TC) for a monopolist is given by

$$TC = 500 + 20Q^2$$

Let the demand equation be given by

$$P = 400 - 20Q$$

Because total revenue is price times quantity, the total revenue equation is

$$TR = 400Q - 20Q^2$$

What are the profit-maximizing price and quantity?

Solution

Calculation Approach. The equations can be used to compute total cost and total revenue at various rates of output. In turn, these data are used to determine marginal cost and marginal revenue. The data for output rates from 1 to 11 are as follows:

Quantity	*Total Cost*	*Total Revenue*	*Marginal Cost*	*Marginal Revenue*	*Profit*
1	$ 520	$ 380	—	—	$−140
2	580	720	$ 60	$340	140
3	680	1,020	100	300	340
4	820	1,280	140	260	460
5	1,000	1,500	180	220	500
6	1,220	1,680	220	180	460
7	1,480	1,820	260	140	340
8	1,780	1,920	300	100	140
9	2,120	1,980	340	60	−140
10	2,500	2,000	380	20	−500
11	2,920	1,980	420	−20	−940

Note that marginal revenue exceeds marginal costs for the first five units, but that the marginal cost of the sixth unit (220) is greater than its marginal revenue. Hence, profits will be maximized by producing five units of output. This result is verified by the profit data in the last column. Producing four units, total profit is $460, while total profit is $500 for five units. However, total profit is only $460 if the sixth unit is produced.

Mathematical Approach. The equation for marginal revenue is the derivative of the total revenue equation with respect to Q. Similarly, marginal cost is the derivative of total cost with respect to quantity. That is:

$$MR = \frac{dTR}{dQ} = 400 - 40Q$$

and

$$\frac{dTC}{dQ} = MC = 40Q$$

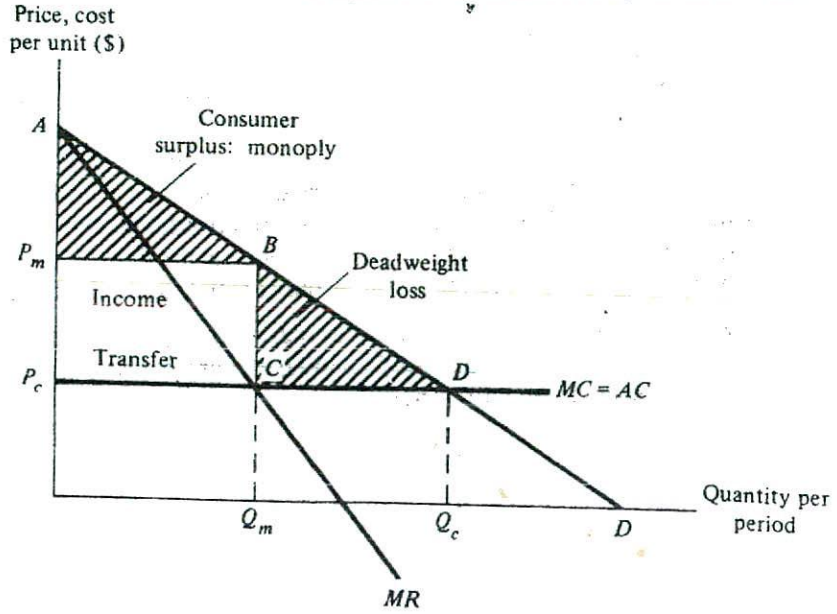
Profits are maximized by choosing the quantity where marginal revenue equals marginal cost. Thus

$$400 - 40Q = 40Q$$

Solving for Q gives five units as the profit-maximizing quantity. Substituting $Q = 5$ into the demand equation gives $P = \$300$.

Allocative Inefficiency and Income Redistribution

Assume that average costs and marginal costs are constant for all output levels, as shown in Figure 9.7. To maximize profit, the firm would equate marginal revenue to marginal cost, produce Q_m, and charge P_m. As an alternative, suppose that policymakers required the monopolist to use the competitive rule of equating price to marginal cost. In that case, the price would be P_c. Those consumers who value the product in excess of P_c would be purchasers, resulting in total sales of Q_c. Because all consumers are charged the same price, most buyers would receive a consumer surplus as a result of their purchase. The dollar value of this surplus is the difference between their valuation of the product (as depicted by the demand curve) and the price, P_c. For example, in Figure 9.7, the person who values the product most highly receives a consumer surplus equal to the vertical distance AP_c. But the consumer who purchases the Q_cth unit receives no surplus because he or she attaches a value to the product just equal to the purchase price. Therefore, when output is determined based on the price equals marginal cost rule, the total consumer surplus is the area under the demand curve and above the price. In the figure this area is the triangle ADP_c.



Now consider output and price of the profit-maximizing monopolist. As indicated, the price will be P_m and the quantity will be Q_m. Although the price is higher than using the competitive pricing rule, there is still an area of consumer surplus created by those consumers who value the product above its price. The consumer surplus in the monopoly case is the area of the triangle ABP_m. This triangle is part of the consumer surplus under competition, ADP_c. The rectangle P_mBCP_c was also a part of the consumer surplus under competition, but now is economic profit earned by the monopolist. This economic profit represents a redistribution of income from consumers to producers. Whether or not this change is considered an improvement requires an assumption about the appropriate distribution of income and cannot be evaluated using efficiency criteria.

Finally, the last component of the consumer surplus under competition is the triangle BDC. This area is referred to as the allocative inefficiency, or *deadweight loss,* associated with monopoly. It represents the loss of consumer surplus stemming from monopoly pricing and is a net loss to society. No assumptions about the relative merits of consumers and producers or the distribution of income are required to assess this impact of monopoly. It is a loss suffered by consumers that is not captured by anyone.

The source of the allocative inefficiency can be identified using Figure 9.7. The monopolist produces until marginal revenue equals marginal cost. Expanding production beyond Q_m would result in reduced profit because the incremental revenues are less than the extra costs incurred. However, at the output rate, Q_m, the consumer is charged a price, P_m, that is greater than marginal cost. Thus, only those consumers who attach a value to the product of at least P_m will purchase if price is set at that level.

Note that the last person to buy values the product at exactly P_m, but the cost of producing is given by the marginal cost curve and equals P_c. Hence, the value to the last buyer is greater than the opportunity cost of production. Thus, social welfare would be increased if more were produced. Specifically, expanding production by one unit would generate additional consumer surplus equal to the vertical distance between the demand curve and the marginal cost curve. Additional increases in output would create successively smaller consumer surplus gains until the output rate Q_c was reached. At that point there would be no additional consumer surplus. Thus, the triangle BDC can be thought of as the loss of consumer surplus stemming from the output-restricting tendency of monopoly.

To summarize, relative to firms in perfect competition, monopolists produce too little output and set too high a price. Whereas competition results in the lowest price consistent with the survival of the firm, the monopolist charges the highest price consistent with profit maximization. From the perspective of society, resource allocation would be improved if more resources were used to produce the products provided by the monopolist. There is also an income distribution effect associated with monopoly, but an evaluation of this transfer depends on judgments regarding the relative needs of consumers and producers. Although economic analysis provides little assistance in judging this income transfer, its importance in public policymaking should not be underestimated. In political debate, a legislator's call to take action against the abuses of monopoly is not commonly based on esoteric notions of allocative inefficiency as discussed in a textbook. Rather, it is more likely to focus on the alleged unfairness of the income redistribution from consumers to owners of the monopoly.

Key Concepts

- Monopoly pricing results in allocative inefficiency because not enough output is produced.
- Monopoly pricing also causes a redistribution of income from consumers to the owners of the monopoly.

Case Study

The Price of Caviar and the Fall of Communism

Many products produced in the now-defunct USSR were considered inferior by Western standards. However, one area where the Soviets excelled was in the production of caviar. In the Volga River, near the Caspian Sea, the water temperature and degree of salinity are a perfect spawning ground for the sturgeon whose eggs produce the world's most prized caviar.

During the nearly seven decades of communist rule, the Soviet state maintained a near monopoly over the harvest, processing, and marketing of this delicacy. The result was a textbook example of the restricted output, high prices, and income redistribution associated with monopoly control over a market.

Until 1991, the Soviet Bureau of Fisheries made virtually all decisions about sales of Russian caviar. In a typical year, about 2,000 tons of caviar were harvested. Of this amount, the Bureau allowed only 150 tons to be exported. By restricting the amount available to foreign consumers, the price was maintained at an extremely high level. For example, in Moscow, the black-market price for top-grade black caviar in 1991 was about $5 per kilogram. But the same caviar could easily have been sold for $500 to $1,000 per kilogram in New York City. Clearly, the monopoly arrangement caused a substantial redistribution of income from New York restaurant and delicatessen patrons to the Soviet state. During the period of communist rule, caviar was a much-needed source of hard currency for the government.

One effect of the breakup of the USSR in 1991 was to increase competition in the caviar market. The two largest Soviet fisheries are now under the control of two different republics, Russia and Kazakhstan. In addition, fishermen on the Caspian Sea have begun to bypass the government and establish their own export businesses. The results were as predicted by economic theory. Prices dropped by 20 percent in 1 year. More recently, over-fishing has become a problem and the number of sturgeon have greatly decreased in the region. ■

Technical Inefficiency and Rent Seeking

There are at least two other negative consequences of monopoly. The first has come to be known as *technical inefficiency.* In discussing both perfect competition and monop-

oly in this chapter, it was assumed that the goal of managers is to maximize profit. A necessary condition for profit maximization is that costs be minimized for the output rate selected. Consequently, all cost curves shown in the figures represent minimum cost production. For firms in a competitive market, this is a reasonable assumption because the manager may have no choice. In the long run, prices are driven down until perfectly competitive firms earn only a normal rate of profit. Hence, businesses that are inefficient will not be able to survive in competition with more efficient rivals because they will earn inadequate profits to sustain their operations. In a competitive market, cost minimization is a necessity.

However, a monopolist may not be under the same constraint. Earning economic profits, the manager of a firm that is insulated from competition has some discretion with respect to cost minimization. If costs increase, the survival of the monopoly firm will not be in jeopardy because the business has a cushion of economic profits. But there is a trade-off. Any increase in costs resulting from waste will reduce economic profits earned by the firm.

Why would a manager sacrifice profits by permitting the firm to operate inefficiently? There are several possible reasons. One is that cost minimization requires effort. The search for least-cost resources and the most advanced technology can be difficult. A manager, particularly one who is salaried and not a stockholder of the firm, may choose to go home a little earlier or take a little longer lunch rather than devote full effort to managing. Another explanation is that firms are run by people, and people are likely to make mistakes. Decisions are made under conditions of uncertainty by individuals of different capabilities who experience stress and illness. Sometimes these choices may not be wise when judged from the perspective of hindsight.

Labor contracts are another source of technical inefficiency. When workers are hired, the agreement usually is concerned with inputs—the number of hours to be worked. A new employee provides his or her time but usually does not commit to a specified level of effort. Just as managers prefer leisure to work, so do employees. A primary task of management is to monitor performance, but this is not a costless task. Consequently, in many organizations, there is considerable slack.

Although managers may permit technical inefficiency, it is not consistent with the objectives of stockholders. Every dollar wasted is a dollar in reduced profit. To the extent that a firm's stockholders can constrain the behavior of managers, the problem of technical inefficiency will be reduced. But in today's economic system, stockholder control is often limited. Most large firms have thousands of stockholders, and any single individual or group does not control a large proportion of the stock. This separation of ownership and control provides opportunities for technical inefficiency where monopoly power exists.

Yet another consequence of monopoly is the tendency for *rent-seeking* behavior. The ability of a monopolist to earn economic profit is a valuable possession. Any rational person should be willing to pay to obtain and maintain this privilege. If a firm earns $1 million in economic profits per year, the value of its monopoly position is worth up to that amount. That is, the owners of the firm would be willing to spend up to $1 million each year to assure a continued flow of profits. When resources are expended to seek or maintain monopoly profits, this behavior is referred to as rent seeking.

Rent-seeking behavior does not increase the amount of goods and services produced. Remember that economic profit represents a transfer of wealth from consumers to stockholders. But rent seeking is an attempt to capture these economic profits. Hence, the resources used do nothing more than alter the distribution of income. Sometimes rent seeking may be directed toward obtaining income from consumers. In other cases, the purpose may be to obtain economic profits currently being earned by another firm.

Rent seeking results in a deadweight loss because there is no new productive activity. The analogy may be a bit overstated, but rent seeking has been likened to the activities of a burglar. The thief expends time and effort to break into a person's house, and the potential victims install burglar alarms and locks to prevent theft. But all these efforts either facilitate or prevent changes in the existing distribution of income. The burglar's purpose is to acquire the possessions of the victim and the victim's intent is to avoid losing personal property. Rent seeking is much the same. Nothing new is created. All of the effort is to change the division of the existing pie.

Rent-seeking behavior can take many forms. Often it involves government officials. Choices of policymakers can significantly affect the distribution of income. The location of a defense base in an area or the award of a contract to a firm can mean millions of dollars to the beneficiaries. Changes in tax laws can increase or decrease profits by huge amounts. The decisions of regulatory bodies on who will be allowed to enter an industry or procedures to be followed in conducting business can have an enormous impact on profitability.

Unlike the Ten Commandments, public policy is not etched in stone. Decisions can be affected by activities of groups and individuals. Legislators are constantly besieged by lobbyists, who argue for favorable legislation. Tens of millions of dollars are spent each year on campaign contributions. A firm may engage in a costly public relations campaign to improve its image with policymakers.

Rent seeking can also have a darker side. Because the stakes are so high, bribes for favorable legislative or agency treatment may be offered. Firms may resort to espionage activities to get ahead of competitors. The legal system may be utilized to harass competitors or to forestall legislation or regulations that could have an adverse effect on the firm.

Waste from rent seeking can actually exceed the total amount of economic profit that could be earned. Although no single firm would spend more than it expects to gain, where there are several contending firms, their efforts may offset one another and exceed the potential prize. The outcome is analogous to shoppers during a sale in a china store. In their attempt to get the best buy, they may break more dishes than they purchase.

Technical inefficiency and rent-seeking behavior imply that the observed rate of profit for monopolists may not be high. As shown in Figure 9.8, economic profits may be absorbed by these two sources of deadweight loss. The rectangle P_mBCP_c represents economic profit that could be earned by a monopolist. But where costs are allowed to increase and resources are used to obtain or maintain a monopoly position, the amount of economic profit could be much smaller, as shown in the figure. In fact, it is possible that inefficiency and rent seeking could consume all the economic profit, and the monopolist would earn only a normal return.

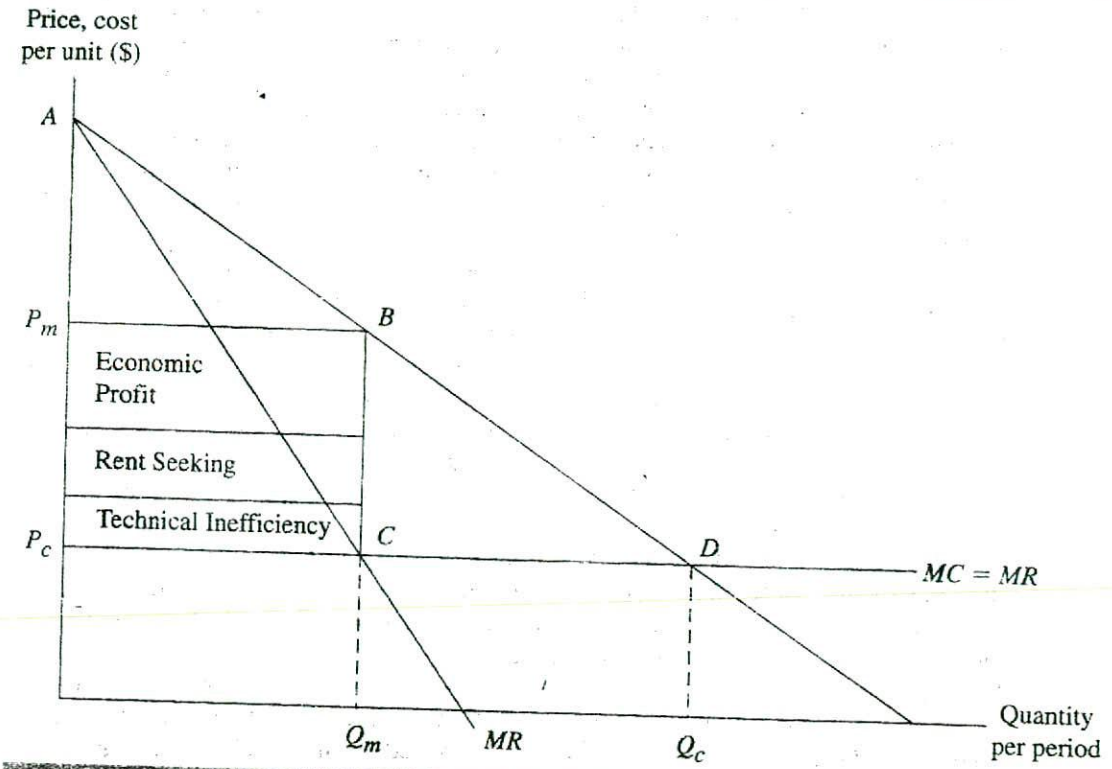


FIGURE 9.8 Effects of Technical Inefficiency and Rent Seeking on Economic Profit

Key Concepts

- If a firm has market power and there is separation of ownership from control, technical inefficiency may result because of the failure of managers to minimize costs.
- Rent seeking involves the use of resources to acquire or maintain monopoly profits. Rent seeking involves a deadweight loss to society because no additional goods or services are produced.

RELEVANCE OF PERFECT COMPETITION AND MONOPOLY

Over 75 years ago, economist Frank Knight wrote: "In view of the fact that practically every business is a partial monopoly, it is remarkable that the theoretical treatment of economics has related so exclusively to complete monopoly and perfect competition."[2] Since that time, important contributions have been made to the analysis of alternative market structures. However, classroom time typically is still heavily weighted toward discussing the extremes of competition and monopoly. What is the rationale for such emphasis?

[2]F. Knight, *Risk, Uncertainty and Profit* (Chicago: University of Chicago Press, 1985), p. 193.

The assumptions of the model of perfect competition are very restrictive. Few markets would meet the requirements of many small sellers, easy entry and exit, and an undifferentiated product. Certain parts of the agricultural sector are the most likely candidates. For example, in the midwestern grain exchanges, there are many buyers and sellers of wheat. Wheat of a given grade is relatively homogeneous. Compared to activities such as electricity production, resources can enter and leave agriculture without great difficulty. However, to claim that resources are perfectly mobile strains credibility. All in all, it is not easy to think of many circumstances where the requirements for perfect competition are closely approximated.

It is also difficult to think of examples of pure monopoly. In most communities the local electric utility usually fits the definition of being the single seller of its product. However, there may be substitutes for their products that reduce the monopoly power of these firms. For example, an electric utility does face competition from substitute energy sources. Consumers usually have the option of heating their homes and water, cooking, and drying clothes with gas instead of electricity. Large industries may generate their own electricity if the rates of the local utility become too high.

If there are few, if any, examples of pure monopoly and perfect competition, why expend the time and effort required to discuss these extreme conditions? In many applications, it is useful to think of perfect competition and pure monopoly as extremes, with other market structures positioned in between. Although there may be few industries at either extreme, there are many that have most of the characteristics of perfect competition or monopoly. Hence, the value of the extreme models is that they serve as benchmarks. Industries that approximate perfect competition are likely to function much like those in the perfectly competitive model. In contrast, those that have many of the characteristics of monopolies will generate monopoly-like results. Indeed, the monopoly model may be applicable in many situations where producers sell a differentiated product and believe that they have some control over price. In addition to providing information on the likely behavior and result of specific market structures, the extreme models of competition and monopoly provide guidance in making public policy. As a general rule, economists favor policies that move industries toward the competitive end of the spectrum.

Key Concepts

- Perfect competition serves as a benchmark and a guide for public policy.
- The monopoly model may be applicable in situations where a producer has some power over price.

SUMMARY

Market structures can be characterized on the basis of four characteristics: (1) number and size distribution of sellers, (2) number and size distribution of buyers, (3) product differentiation, and (4) ease of entry and exit. The model of perfect competition assumes a large number of small buyers and sellers, undifferentiated products, and ease of entry and exit. Firms in a perfectly competitive market face a demand curve that is horizontal at the equilibrium price. This price is determined by the interaction of the market supply and demand curves. Because they have no control over price, the objective of managers is to determine the rate of output that maximizes profit.

The profit-maximizing output for the perfectly competitive firm occurs where price equals marginal cost. In the short run, firms in perfect competition may earn economic profit. But if price drops below average variable cost, the firm should shut down. However, in the long run, entry of new firms and/or plant expansion drive price down and eliminate the economic profit. Perfect competition results in efficient allocation of resources because production occurs at minimum average cost, voluntary exchange is maximized, and capital is employed in its highest value use.

The monopolist is a single seller of a differentiated product. Entry into the market is difficult or prohibited. As the single seller, the monopolist has power over price. The decision rule for maximizing profits is to produce until marginal revenue equals marginal cost and then charge the price that the demand curve will allow. Because entry is restricted, the monopolist may earn economic profits in both the short and the long run.

Consumer surplus is the difference between what consumers are willing to pay for a product and the price that must be paid to purchase it. The principle of consumer surplus can be used to demonstrate that monopoly pricing causes allocative inefficiency because too little of the product is produced. Monopoly pricing also results in a redistribution of income from consumers to the stockholders of the firm.

If a firm has market power and there is a separation of ownership and control, technical inefficiency can also occur because the firm's managers will not be required to minimize costs. When monopolists use resources to acquire or maintain economic profits, this is referred to as rent-seeking behavior. Because no additional goods or services are produced, rent seeking imposes a deadweight loss on society. The combination of technical inefficiency and rent seeking may cause the economic profit earned by a monopolist to not appear excessive.

Few market structures meet the restrictive assumptions for perfect competition or monopoly. Still, these economic models are useful because the performance of many industries approximates the outcomes of perfect competition or monopoly. Also, the perfectly competitive model is often used as a benchmark for evaluating the performance of actual markets and as a guide for public policy.

Discussion Questions

9-1. Why is concrete sold in local markets, while cement powder is sold in a national market?

9-2. Does product differentiation always refer to real differences between products? Use an example to explain your answer.

9-3. Does ownership of very specialized capital equipment affect ease of exit from an industry? Why or why not?

9-4. How would risk affect the normal rate of profit in an industry?

9-5. Suppose that firms in a perfectly competitive industry are earning less than a normal rate of profit. In the long run, what price adjustments will occur in this industry? What will cause these adjustments?

9-6. Basically, perfectly competitive firms and monopolists use the same rule to determine the profit-maximizing output. True or false? Explain.

9-7. Firms in a perfectly competitive market do not have to compete with the other individual firms in the market. True or false? Explain.

9-8. In the long run, firms in a perfectly competitive market produce at the minimum point on their average cost curves. However, the long-run profit-maximizing output for a monopolist will not be at the point of minimum average cost. Does this mean that competitive firms can produce at a lower average cost than the monopolist? Explain.

9-9. How is the deadweight loss from monopoly affected by the slope of the demand curve?

9-10. Do nonprofit institutions, such as universities, ever engage in rent-seeking behavior? Give an example.

Problems

9-1. Suppose the market supply and demand equations for plywood are given by

$$Q_S = 20{,}000 + 30P$$

and

$$Q_D = 40{,}000 - 20P$$

a. Graph the supply-and-demand equations and show the equilibrium price and quantity.

b. Determine the equilibrium price and quantity algebraically.

c. Suppose an increase in housing starts results in a new demand equation,

$$Q_D' = 50{,}000 - 20P$$

What is the new equilibrium price and quantity?

9-2. The market supply-and-demand equations for plywood are the original equations used in problem 9-1. The plywood industry is perfectly competitive, and the marginal cost equation for one firm, High Country Plywood, is given by

$$MC = 200 + 4Q$$

a. What is the short-run profit-maximizing output rate for High Country Plywood?

b. Average cost is given by

$$AC = \frac{1{,}000}{Q} + 200 + 2Q$$

In the short run, how much economic profit will the firm earn?

9-3. Tyson Brothers Manufacturing is currently earning economic profit. However, the market for the firm's product is perfectly competitive, so the economic profit is not expected to persist in the long run. Tyson's total and marginal cost functions are given by

$$TC = 500Q - 20Q^2 + Q^3$$

and

$$MC = 500 - 40Q + 3Q^2$$

a. At what output rate will average costs be a minimum?

b. If the cost curves for all other firms in the market are the same as Tyson's, determine the long-run equilibrium price.

9-4. The equilibrium price in a perfectly competitive market is \$10. The marginal cost function is given by

$$MC = 4 + 0.2Q$$

The firm is presently producing 40 units of output per period. To maximize profit, should the output rate be increased or decreased? Explain.

9-5. The market supply and demand curves for a product are given by

$$Q_s = 3{,}000 + 200P$$

and

$$Q_D = 13{,}500 - 500P$$

The industry supplying the product is perfectly competitive. An individual firm has fixed costs of \$150 per period. Its marginal and average variable cost functions are

$$MC = 15 - 4Q + \frac{3Q^2}{10}$$

and

$$AVC = 15 - 2Q + \frac{Q^2}{10}$$

a. What is the profit-maximizing rate of output for the firm?
b. What is the maximum total profit for the firm?

9-6. United Electric is the sole supplier of electricity to the community of Lakeview. Managers of the firm estimate that the demand for electricity is given by

$$P = 200 - 4Q$$

The firm's marginal cost equation is given by

$$MC = 4Q$$

a. What is the profit-maximizing price and quantity?
b. Can economic profit be determined from the information given? Why or why not?

9-7. A consultant estimates that the demand for the output of Marston Chemical is represented by the equation

$$Q = 2{,}000 - 50P$$

a. If the managers of Marston decide to maximize total revenue instead of profit, at what output rate should the firm operate? What is the revenue-maximizing price?
b. Will the revenue-maximizing output be greater than or less than the profit-maximizing output rate? Explain.

9-8. The demand equation faced by a monopolist for a product is given by

$$Q = 50 - 5P$$

A price of \$5 is charged for the product.
a. Using this information, draw a graph that shows the consumer surplus.

b. Compute the amount of consumer surplus generated by sale of the product.

9-9. Lyon Concrete is a monopoly supplier of concrete in northern Arkansas. Demand for the firm's concrete is given by

$$P = 110 - 4Q$$

Marginal cost is constant and equal to 10.

a. What are the profit-maximizing price and output?
b. What is the deadweight loss resulting from Lyon's monopoly?
c. Compared to pricing at marginal cost, how much income is redistributed from consumers to the owners of the monopoly?

9-10. Show that the profit-maximizing quantity for a monopolist will always lie in the elastic region of the demand curve.

9-11. The demand equation for a monopolist is given by $P = 50 - 2Q$ and the marginal cost is \$10.

a. Compute the deadweight loss associated with monopoly pricing.
b. If $P = 50 - 4Q$, what is the deadweight loss?
c. Based on your answers to (a) and (b), how is the deadweight loss related to the slope of the demand curve?

9-12. The market price faced by a firm in a perfectly competitive market is \$50 and marginal cost is given by $10 + 2Q$.

a. What is the profit-maximizing rate of output?
b. How does a \$1 increase in the price affect the optimal output?
c. How does a \$1 increase in the marginal cost at each output rate affect the optimal output?

Problems Requiring Calculus

9-13. The manager of Biswas Glass Company estimates that total revenue from the sale of her firm's product is given by the equation:

$$TR = 300Q - \frac{Q^2}{2}$$

The total cost equation is estimated to be

$$TC = 5{,}000 + 60Q + Q^2$$

a. What is the profit-maximizing price and output rate? What is the amount of economic profit?
b. At what output rate is average cost a minimum? At this output rate, what is the amount of economic profit?

9-14. For a perfectly competitive firm, the market price is \$16. The total cost equation is

$$TC = \frac{Q^3}{3} - 5Q^2 + 40Q$$

Use calculus to determine the profit-*maximizing* and the profit-*minimizing* rates of output. Explain.

9-15. Michelle's Mints is a small chain of candy stores. Cross-section data from the stores were used to estimate the demand equation. Holding income and prices of other goods constant, the demand equation is estimated to be

$$P = 12Q^{-1/3}$$

where P is price per pound and Q is pounds sold per day per store. The marginal cost of supplying the candy is constant and equal to \$2 per pound.

a. What is the point price elasticity of demand?
b. What are the profit-maximizing price and quantity?

9-16. The manager of a small candy shop operating in a perfectly competitive market determines that his average cost for chocolates is given by $10 - .2Q + .005Q^2$, where Q is in pounds per month. The market price of chocolates is expected to remain at \$8 per pound. How many pounds per month should he produce? Explain.

9-17. The plant manager of a firm producing a specialized brand of caviar believes that her total revenues are given by $TR = 1000Q - Q^2$ and that total costs are $TC = -200Q - Q^2 + Q^3$. What are the profit-maximizing price and quantity for the firm?

9-18. A firm in a perfectly competitive industry (which is in long-run equilibrium) is charging \$5 for its product. The average cost equation is $AC = 50 - 6Q + 0.2Q^2$.

a. How much economic profit is the firm earning? Explain.
b. How many units of output is the firm producing?
c. What is the firm's average cost?

9-19. A monopolist's demand function is given by $P = 80 - 4Q$. The firm's total cost is given by $TC = 10Q + Q^2$.

a. What is the profit-maximizing price and quantity?
b. How much economic profit can the firm earn?

9-20. A local lawn care company is operating in a perfectly competitive environment and estimates that its short-run total costs are given by $TC = 1000Q - 30Q^2 + Q^3$, where Q is the number of lawns cared for. Below what price should the firm shut down its operations?

9-21. A monopolist has total revenue given by $TR = 480Q - 8Q^2$ and total costs given by $TC = 400 + 8Q^2$.

a. What are the profit-maximizing price and quantity?
b. What would the profit-maximizing price and quantity be if the firm made its output decision using the decision rule employed by firms in a perfectly competitive market structure?

CHAPTER 10

Monopolistic Competition, Oligopoly, and Barriers to Entry

- **Preview**
- **Monopolistic Competition**
  - Characteristics
  - Profit-Maximizing Price and Output in the Short Run
  - Profit-Maximizing Price and Output in the Long Run
  - Evaluation of Monopolistic Competition
- **Oligopoly**
  - Characteristics
  - Price Rigidity: The Kinked Demand Model
  - Interdependence: The Cournot Model
  - Cartels and Collusion
  - Price Leadership
- **Market Structure and Barriers to Entry**
  - Sources of Barriers to Entry
  - Spectrum of Market Structures
- **Advertising**
- **Summary**
- **Discussion Questions**
- **Problems**

PREVIEW

Chapter 9 developed models of price and output determination for two important types of market structures: monopoly and perfect competition. But most markets have neither the single seller required to meet the definition of a monopolist nor the large number of small sellers and undifferentiated product necessary to qualify as perfectly competitive. Where the number of sellers is large and the product differentiated, the model of monopolistic competition is a useful tool for analyzing price and output decisions. When there are only a few sellers, oligopoly theory can provide important insights for decision making.

The first section of this chapter develops the model of monopolistic competition. The second considers oligopoly theory in its various forms. In the third section, the relationship between market structure and barriers to entry is examined. Finally, there is a brief discussion of advertising decisions.

MONOPOLISTIC COMPETITION

The models of perfect competition and monopoly are useful, but there is a need to bridge the gap between these extreme forms of market structure. An important contribution is the model of monopolistic competition developed by Edward Chamberlin.[1] Chamberlin observed that even in markets with a large number of sellers, the products of individual firms are rarely homogeneous. For example, consider men's shoes. In a large city there may be hundreds of shoe stores. But men's shoes may be highly differentiated in the minds of consumers. This product differentiation may reflect materials and workmanship of the shoes sold in a particular store, or it may be the result of effective advertising. The manner in which the store displays the shoes can be another source of product differentiation. An establishment with thick carpet and soft music may have an advantage over a firm that stocks its merchandise on shelves like a warehouse. Location is another source of product differentiation. Sellers in a nearby mall will be more likely to obtain a consumer's business than stores on the other side of town.

Characteristics

The theory of monopolistic competition has elements of both monopoly and perfect competition. Like perfect competition, it assumes that there are a large number of small sellers. Thus, the actions of any single seller do not have a significant effect on other sellers in the market. Also, like perfect competition, it is assumed that there are many buyers and that resources can easily be transferred into and out of the industry. However, the model of monopolistic competition resembles the monopoly models in that products of individual firms are considered to be slightly differentiated. That is, the product of one firm is assumed to be a close, but not a perfect, substitute for that of other firms. The result is that each firm faces a demand curve with a slight downward slope, implying that the individual firm has some control over price. Although increasing its price

[1] E. Chamberlin, *The Theory of Monopolistic Competition* (Cambridge, MA.: Harvard University Press, 1962).

TABLE 10.1 Market Structure Characteristics of Monopolistic Competition

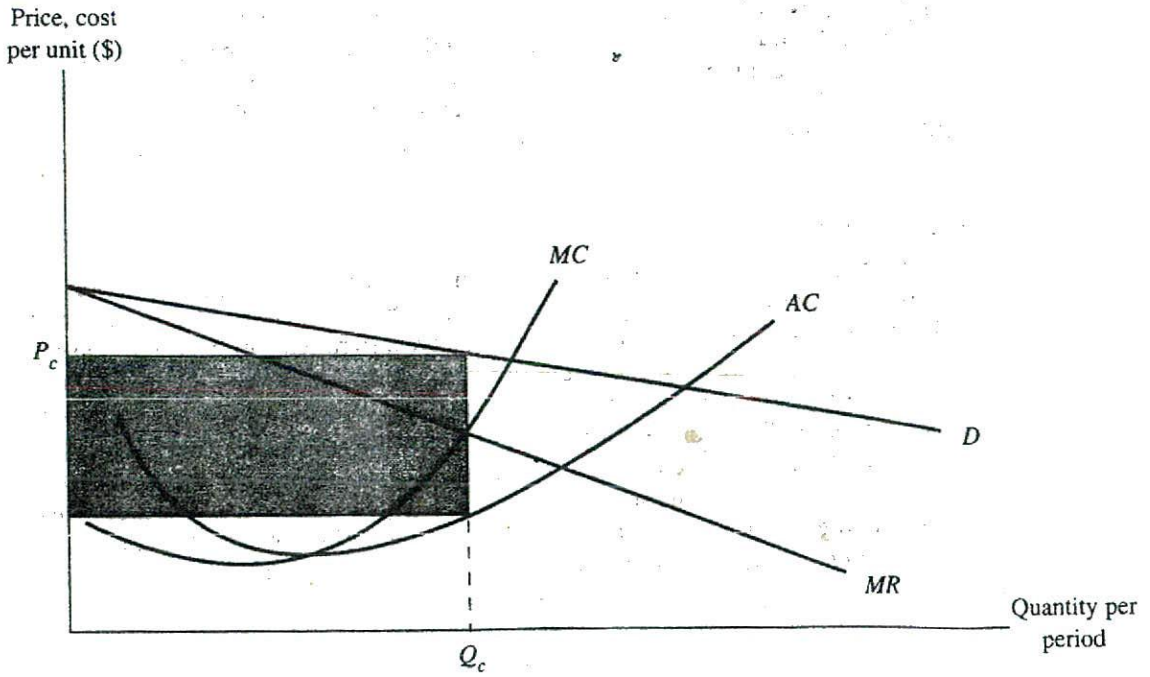
Number and size distribution of sellers	Many small sellers. Actions of individual sellers go unheeded by other firms.
Number and size distribution of buyers	Many small buyers.
Product differentiation	Slightly differentiated. Product of one firm is a fairly close substitute for that of other sellers.
Conditions of early and exit	Easy entry and exit.

will cause the firm to lose sales, some consumers will be willing to buy at the higher price because the product is slightly differentiated from that of competitors. The characteristics of monopolistic competition are summarized in Table 10.1.

Profit-Maximizing Price and Output in the Short Run

Managers of firms in monopolistic competition determine the rate of output, product attributes, and advertising expenditure that maximizes profits. To simplify the discussion, it is assumed that advertising and product attributes have already been determined. Therefore, determining the profit-maximizing rate of output and price are the remaining decisions for managers. Chamberlin's monopolistic competition model also assumes that all firms have similar demand and cost curves. Thus it is possible to consider a "representative" or "typical" firm. The demand, marginal revenue, and cost curves for such a firm are shown in Figure 10.1.

FIGURE 10.1 Short-Run Profit Maximization in Monopolistic Competition



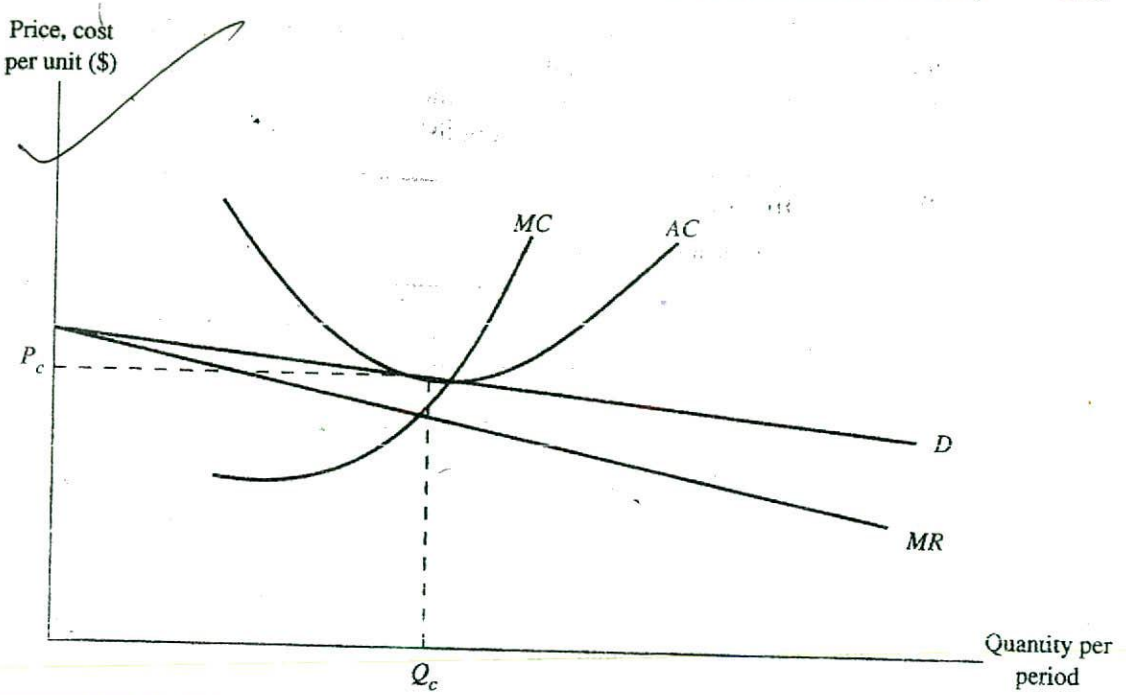


FIGURE 10.2 Long-Run Profit Maximization in Monopolistic Competition

The results of monopolistic competition in the short run are similar to those of monopoly. The profit-maximizing output rate occurs at Q_c, where marginal revenue equals marginal cost. The corresponding price (as determined by the demand curve, D) is P_c. Like a monopolist, a firm in monopolistic competition may earn short-run economic profit. Recall that economic profit per unit is price minus average cost. Thus, total economic profit is shown by the shaded area in Figure 10.1.

Profit-Maximizing Price and Output in the Long Run

In the long run, monopolistic competition generates results similar to those of perfect competition. Because entry into the industry is easy, economic profit induces other firms to enter the market. For example, the success of a video rental outlet in a community may entice other firms to provide this service. Because an inventory of movies is the major prerequisite for entry, new suppliers may include bookstores, gas stations, grocery stores, or other businesses. As this entry occurs, the market shares of existing firms decrease. Thus, the demand curve faced by these firms shifts down and to the left, until it becomes as shown in Figure 10.2.

As in the short run, the representative firm maximizes profit by equating marginal revenue and marginal cost. In Figure 10.2, profit maximization requires setting price at P_c and producing Q_c units of output per period. Note that the demand and average cost curves are tangent at this price–quantity combination. This implies that

price and average cost are equal and so there is no economic profit. Thus there is no incentive for new firms to enter the market. Similarly, because the representative firm is earning a normal return, firms will not exit the market. Hence P_c and Q_c represent the long-run equilibrium for firms in monopolistic competition.

Example **In the Long Run They're All Dead**

DOGGONE, a pet mortuary in Chicago, offers complete funerals for dogs. The pet funeral business in Chicago is monopolistically competitive. The manager of DOGGONE has determined that the firm's demand equation is given by $P = 309.75 - Q$ and the long-run total cost equation is $TC = 400Q - 20Q^2 + Q^3$, where Q is the number of funerals per month.

What is the long-run equilibrium price and quantity and how much economic profit will the firm earn?

Solution In monopolistic competition in the long run, price will be driven down to average cost. The average cost equation is computed by dividing total cost by quantity. Thus, the optimal quantity is the solution to $P = AC$, or

$$P = 309.75 - Q = 400 - 20Q + Q^2 = AC$$

Rearranging terms yields the quadratic equation

$$Q^2 - 19Q + 90.25 = 0$$

which has the single solution, $Q = 9.5$. Substituting this quantity into the demand equation gives $P = 309.75 - 9.50 = \$300.25$. Because price equals average cost, economic profit is zero.

Figure 10.2 on page 343 shows that, in monopolistic competition, marginal revenue must equal marginal cost at the optimal output. For the demand and total cost equations of this problem, $MR = 309.75 - 2Q$ and $MC = 400 - 40Q + 3Q^2$. Substituting $Q = 9.5$ into these two equations gives $MR = MC = 290.75$. Hence, this condition also is fulfilled.

Evaluation of Monopolistic Competition

It is sometimes suggested that firms in monopolistic competition are inefficient. Figure 10.2 shows that the profit-maximizing output does not occur at the minimum point on the firm's average cost curve. Thus, it can be argued that the firm is operating at an inefficient output rate. In contrast, in chapter 9 it was demonstrated that the long-run equilibrium rate of output in perfect competition occurs at the point of minimum average cost.

The difference in the two outcomes is the result of the downward-sloping demand curve in monopolistic competition. Remember, the long-run equilibrium is the point of tangency of the demand and average cost curves. But because the demand curve is not horizontal, the tangency point cannot be at minimum average cost. However, this result does not necessarily imply inefficiency. The downward slope of the demand curve is the result of product differentiation in the market. Presumably, these differences are of value to consumers as they select goods that meet their particular needs. For example, although name-brand canned goods are likely to cost more than generic brands, many buyers are willing to pay the extra price as an assurance of quality. In general, the va-

lidity of the claim that monopolistic competition is inefficient depends on a comparison of the benefits derived from product differentiation and the increased costs caused by differentiated products.

Key Concepts

- Firms in monopolistic competition have some control over price because their products are differentiated.
- As with perfect competition, there may be economic profits in the short run, but there are no long-run economic profits in monopolistic competition.

Case Study

Competition in the Video Rental Industry

A good argument can be made that today's market for the rental of videocassettes is monopolistically competitive. Even in medium-sized cities, there usually are many places where a movie can be rented. Some of these outlets focus exclusively on video rentals, but many music stores, gas stations, and grocery stores also have movies available for rent. In fact, the typical video outlet in a metropolitan area in the United States now has six competitors within a 3-mile radius. Although there are some large firms in the industry, market concentration is relatively low—chains with as many as 50 outlets have only a 15 percent share of the total market.

Product differentiation is also a characteristic of the video rental industry. While it is true that a particular movie from one store is identical to the same movie from another outlet, sellers can differentiate themselves in other ways. A music store may offer tapes and CDs in addition to movies, while a grocery store provides the opportunity to pick up a movie along with needed food items. An outlet that only rents videocassettes may have the advantage of maintaining a huge selection of movies.

The video industry also meets the entry and exit requirements for monopolistic competition. The basic requirement to become a participant is to have display space, an inventory of tapes, and a computer system for record keeping. New movies are easily available from distributors, and there is a rather active market for the inventories of firms who wish to exit the market.

The history of the video rental industry is consistent with the predictions of the model of monopolistic competition. During the early 1980s, there were far fewer outlets than today. Because of the lack of competition, tapes rented for as much as $8 per day, and the business could be extremely profitable. Some early entrants had profits as high as 80 percent of sales. But entry occurred rapidly, and prices dropped precipitously. Today the prices of new releases are as low as $1.49, and older movies can be rented for less than a dollar a day. The decline in prices has also affected profits. Video rental outlets in the 1990s are fortunate if profits are 10 percent of sales, and many firms have

been forced to leave the market because of their losses. Because competition has eliminated economic profits, the rate of entry into the industry is much less than in previous years. In the future, rental outlets will face additional competition from cable television, which has begun to implement the technology to give viewers the opportunity to select a movie of their choice from the convenience of their own homes. ■

OLIGOPOLY

The term *oligopoly* comes from the Greek words *oligos* and *polis* and means, literally, few sellers. Oligopoly is a common form of market structure in modern economic systems. The cereal, automobile, and steel industries in the United States would all qualify. However, oligopolies exist at the local as well as the national level. For example, although there are thousands of movie theaters scattered throughout the nation, the typical consumer considers only a few nearby locations. Other theaters that are farther away may offer lower prices or better food, but proximity is probably the dominant consideration. Hence, the market for movies faced by the individual consumer could be described as an oligopoly.

Characteristics

An oligopoly involves an unspecified number of buyers but only a small number of sellers. There is no precise limit on the number of sellers that a market can have and still be characterized as an oligopoly. The key issue is not numbers, but rather the reaction of sellers to one another. In the three forms of market structure described thus far, there was no need for sellers to be concerned about the actions of individual competitors. The monopolist has no rivals, while firms in perfect and monopolistic competitive markets are too small to have a significant impact on other firms. In contrast, the actions of each firm in an oligopoly do affect the other sellers in the market. Price cutting by one firm will reduce the market share of other firms. Similarly, clever advertising or a new product line may increase sales at the expense of other sellers.

Figure 10.3 shows how the actions of rivals can affect the demand curve faced by oligopolists. If one firm reduces its price and the other firms in the market do not respond, the price cutter may substantially increase its sales. This result is depicted by the relatively elastic demand curve, *dd*. For example, a price decrease from P_1 to P_2 will result in a movement along *dd* and increase sales from Q_1 to Q_2 as customers take advantage of the lower price and abandon other suppliers. However, if the price cut is matched by other firms, the increase in sales will be less. Since other firms are selling at the same price, any additional sales must result from increased demand for the product. Thus, the effect of the price reduction is a movement down the relatively less elastic demand curve, *DD*. Now the price reduction from P_1 to P_2 only increases sales to Q'_2.

Clearly, the responses of competitors can have a significant impact on the outcome of managerial decisions in an oligopoly market. Consequently, decision making in an oligopoly is much more difficult than in other market situations.

The other two characteristics that categorize market structure are product differentiation and condition of entry and exit. The product sold in an oligopoly can be ho-

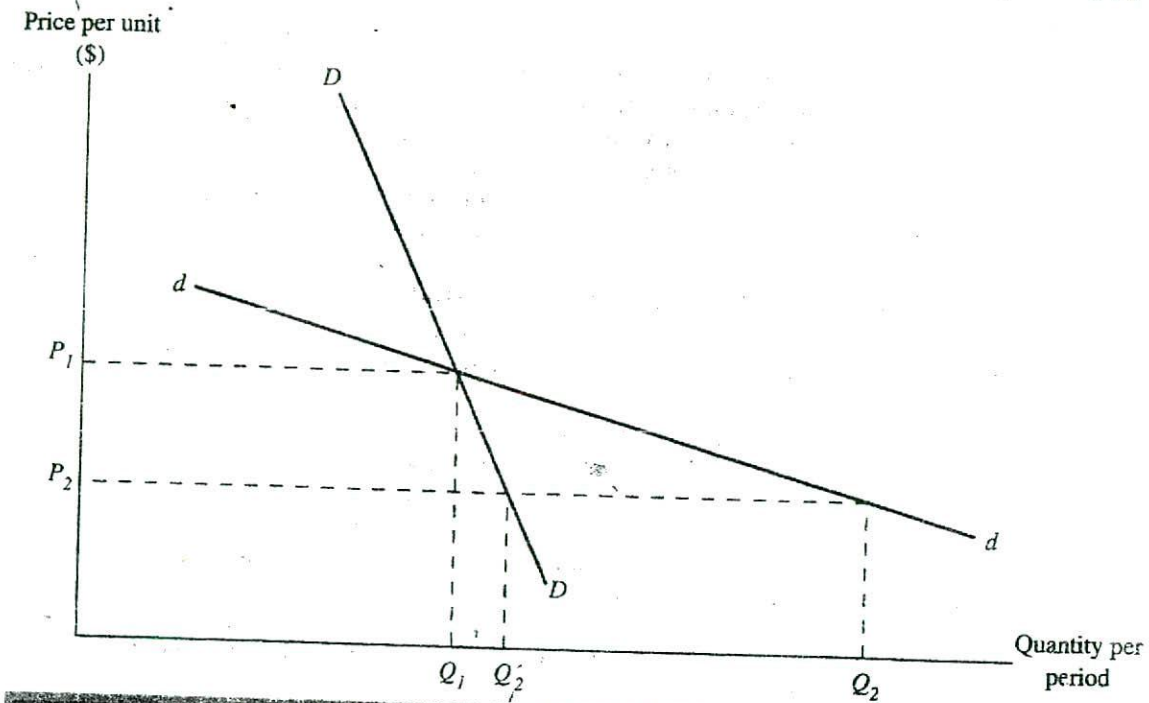


FIGURE 10.8 Demand Curves for an Oligopolist

mogeneous or differentiated. If the product is homogeneous, the market is said to be a pure oligopoly. The steel and copper markets in the United States would fit into this category. If the product is not homogeneous, the market is a differentiated oligopoly. The automobile and television industries are examples of differentiated oligopolies.

With respect to condition of entry, for an oligopolistic market structure to persist in the long run, there must be some factor that prevents new firms from entering the industry. For example, a drug manufacturer might hold a patent that legally prevents other firms from producing the drug covered by the patent. Market structure characteristics of oligopoly are summarized in Table 10.2.

The most distinctive feature of an oligopolistic industry is that sellers must recognize their interdependence. That is, the action of one seller may affect another and, thus, cause that seller to respond in ways that will affect the first seller. Oligopolists are likely to deal with this interdependence in different ways, depending on the specific nature of the industry. In some cases, most actions of competitors will be ignored. In other situations, a

Number and size distribution of sellers	Small number of sellers. Each firm must consider the effect of its actions on other firms.
Number and size distribution of buyers	Unspecified.
Product differentiation	Product may be either homogeneous or differentiated.
Conditions of entry and exit	Entry difficult.

price war may occur in response to a seemingly innocuous price change. Many factors, such as industry maturity, nature of the product, and methods of doing business, can affect the way firms respond to actions of rivals. The difficulty of formulating models of oligopoly stems from the many ways that firms interact. Consequently, there is no general model of oligopoly. There are, however, models that analyze oligopoly decisions on the basis of specific assumptions about the interaction between firms. Several models that reflect specific aspects of oligopolistic interdependence are discussed in this section.

Price Rigidity: The Kinked Demand Model

Early students of oligopoly noted that prices in some markets sometimes remained unchanged for long periods of time. For example, the price of steel rail was set at $28 per ton in 1901 and did not change for 15 years. Between 1922 and 1933, the price remained at $43. Similarly, sulfur prices hovered within a few cents of $18 per ton between 1926 and 1938. This apparent price rigidity led Paul Sweezy to suggest that oligopolists behave as if facing a kinked demand curve.[2] Such a curve is shown in Figure 10.4.

The kink in the demand curve stems from an asymmetry in the response of other firms to one firm's price change. Suppose that the price initially is at P_k, the point of the kink in the demand curve. Sweezy argued that if one firm raised its price, other sellers might not follow the increase. The result would be that the firm would lose a significant amount of sales. This is shown in Figure 10.4 as a relatively elastic demand curve above the existing price, P_k.

In contrast, if the firm reduces its price below P_k, it is likely that the other firms will follow suit in an attempt to maintain their market shares. As a result, the price cut by the original firm will not add much to its sales. Figure 10.4 depicts this outcome as a relatively inelastic demand curve below P_k.

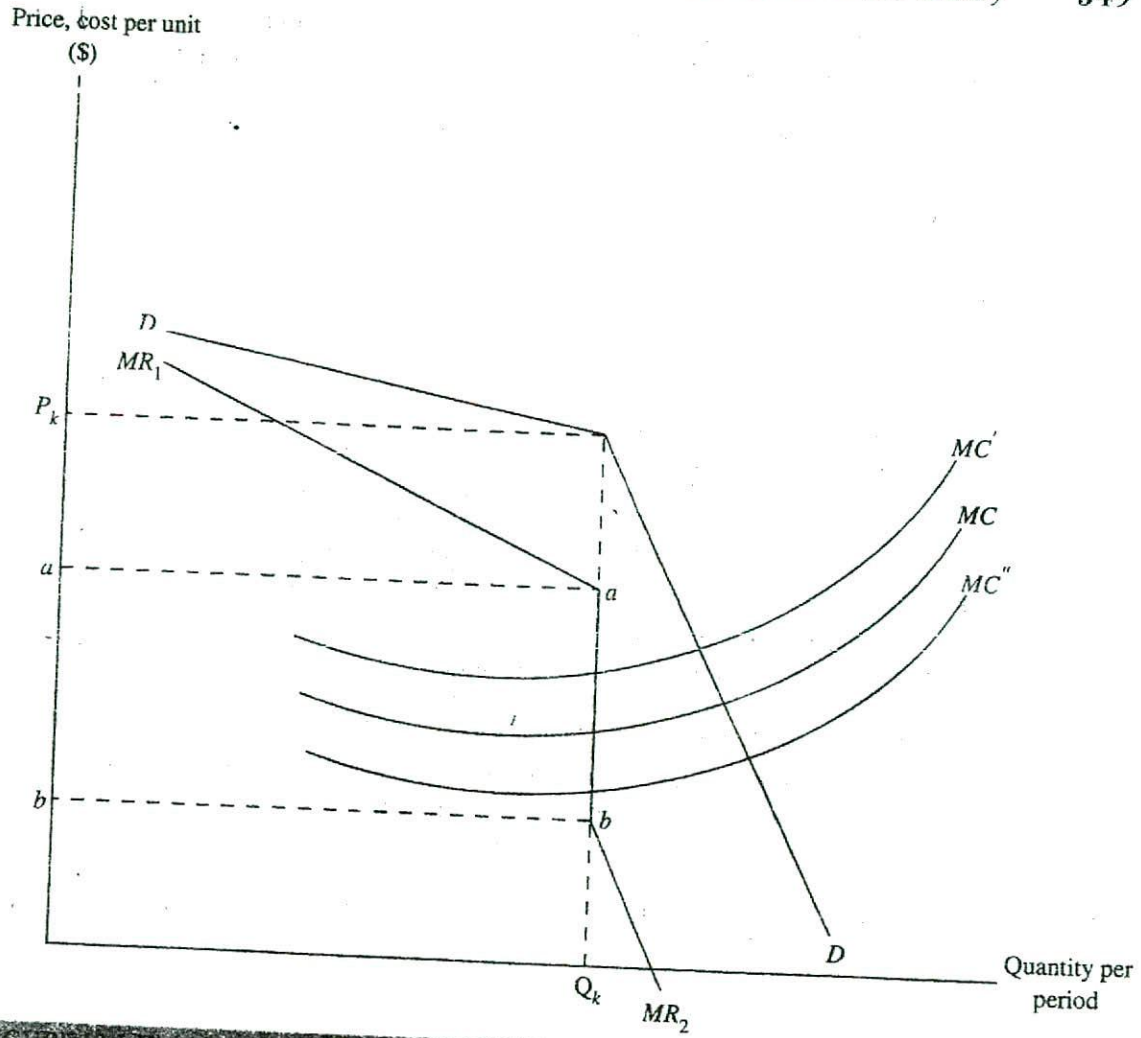
Associated with the demand curve is a marginal revenue curve. For a linear demand curve, the absolute value of the slope of the corresponding marginal revenue is twice as great. Note that the kinked demand curve of Figure 10.4 consists of two linear curves joined at P_k. Marginal revenue for prices above the kink is given by MR_1. For those below the kink, it is MR_2. At the point of the kink, the marginal revenue curve is a vertical line that connects the two segments.[3]

The model assumes that the firm has a U-shaped marginal cost curve, MC, as shown in Figure 10.4. As with previous models, the profit-maximizing output is determined by equating marginal cost and marginal revenue. This output rate is Q_k, and the price (as given by the demand curve at Q_k) is P_k. Note that the profit-maximizing solution occurs at the kink because it is in the region where marginal revenue and marginal cost intersect.

Now suppose that increases in input prices cause the marginal cost curve to shift upward to MC'. Profit maximization requires that marginal revenues again be set equal to marginal costs. But for the marginal cost curve MC', the optimal output is still Q_k, and the optimal price is still P_k. Although the marginal cost has increased, there is no change in the profit-maximizing price and quantity. The explanation is the vertical section of the marginal revenue curve found at the kink in the demand curve. Even though

[2] P. Sweezy, "Demand Conditions Under Oligopoly," *Journal of Political Economy* (August 1939): 568–573.

[3] Because the demand curve of Figure 10.4 has a kink at Q_k, the marginal revenue curve is said to be discontinuous at Q_k. The vertical line, *ab*, represents this discontinuity.



the marginal cost curve shifted upward, it still intersects marginal revenue in the region where that curve is vertical. Hence, there is no change in the optimal output rate and price in response to the input price increase.

Similarly, assume that more efficient production techniques or lower input prices allowed the marginal revenue curve to shift downward, as shown by MC'' in Figure 10.4. As long as the new marginal cost curve intersects the vertical portion of the marginal revenue curve, there will be no change in the profit-maximizing price and quantity. The firm will continue to produce Q_k units and the price will remain at P_k. For price and quantity to change, the marginal cost curve must shift enough to cause it to intersect the marginal revenue either above point *a* or below point *b*.

The important implication of the kinked demand curve model is that firms in oligopolistic market structures could experience substantial shifts in marginal costs and still not vary their prices. This theoretical result is consistent with Sweezy's observation that some oligopolistic markets exhibit very stable prices.

Case Study
Something's Rigid in Denmark

An interesting and well-documented example of price rigidity involved a leather tannery in Denmark. While interviewing the managers of Danish firms about their pricing policies, an economist discovered a firm that charged a higher price for dyed shoe leather than it did for black leather. The price differential had existed since 1890, when dyed leather was more expensive to make. However, by the time of the interview, the dyed shoe leather had become less costly to produce. When queried as to why the pricing policy had not been changed, the firm's manager responded:

> Perhaps we ought to raise the price of black leather somewhat and lower the price of dyed leather to a corresponding degree, but we dare not do so. The fact is that we shall then run the risk of being unable to sell black leather shoes, whereas our competitors will also reduce their prices for dyed shoes.*

The manager's explanation is consistent with the kinked demand curve model. An increase in the price of black leather shoes, unmatched by competitors, was perceived as resulting in a substantial loss of sales. In contrast, a price cut on dyed shoes was expected to result in a price cut by competitors and, hence, little increase in sales. Thus, the price differential between black and dyed shoes was maintained even though relative costs had changed. ■

*B. Fog, *Industrial Pricing Policies* (Amsterdam: North-Holland, 1960), p. 130.

The kinked demand curve model can be criticized on at least two grounds. First, although it explains the reluctance of oligopolists to change prices, it provides no insight into understanding how the price was originally determined. Thus the model is incomplete at best. Second, empirical research has not verified the predictions of the model. George Stigler studied pricing in seven oligopolies.[4] He found that firms in these industries were just as likely to match a price increase by a competitor as they were to follow a rival's cut in price.

[4]G. J. Stigler, "The Kinky Oligopoly Demand Curve and Rigid Prices," *Journal of Political Economy* (October 1947): 432–449.

When first proposed, Sweezy's kinked demand curve model was hailed as a general theory of oligopoly. Today, it is viewed in a more limited perspective as one of several descriptions of oligopoly behavior. In markets where the important firms are of nearly equal size, the product is homogeneous, and sellers are not yet certain how rivals will react, the kinked demand curve model can be a useful tool for understanding pricing particles. Where these conditions are not met, this approach is less useful.

Key Concepts

- The kinked demand curve model of oligopoly is based on the assumption that rivals will match price reductions, but not price increases.
- The kinked demand curve model predicts that price changes will be infrequent in oligopolistic markets.

Interdependence: The Cournot Model

The distinctive feature of the different oligopoly models is the way they attempt to capture the interdependence of firms in the market. Perhaps the best known is the Cournot model, which was developed by a French mathematician in the early 1800s. Although the basic model is rather simplistic, it provides useful insights into industries with a small number of sellers.

Cournot Duopoly Consider a product for which demand is given by the equation $P = 950 - Q_T$, where Q_T is the total amount produced by all of the suppliers in the market. Assume that the marginal and average costs are constant and equal to \$50. The assumption of constant costs is not critical to the analysis, but it simplifies the calculations and makes the insights of the model more obvious.

As a starting point, think about the results of having a monopolist in this market. A single firm would select its output by equating marginal revenue and marginal cost. For the demand equation given, the corresponding marginal revenue equation is $MR = 950 - 2Q_T$. Thus, the profit-maximizing quantity is the solution to

$$950 - 2Q_T = 50$$

or $Q_T = 450$. Substituting this rate of output back into the demand equation gives $P = \$500$.

Next consider a perfectly competitive market. In this situation, prices are driven to costs. Hence, the optimal quantity is determined by $P = MC$, which is the solution to

$$950 - Q_T = 50$$

or $Q_T = 900$. The corresponding price is $P = \$50$. As predicted by economic theory, price is higher and the rate of output lower for a monopoly supplier than in a perfectly competitive market.

Now consider a market that has two sellers—a duopoly. In analyzing this case, an assumption must be made regarding how the two firms respond to one another. The Cournot model assumes that each firm chooses a rate of output to maximize its profits, in the belief that the other firm will continue to produce the same rate of output as it did in the previous period. Although each firm will, in all probability, change its output from period to period, the two firms are assumed to remain oblivious to this adjustment.

For the duopoly case, Q_T in the demand equation $P = 950 - Q_T$ is the sum of the output produced by the first firm, q_1, and the second firm, q_2. Because firm 1 believes that firm 2 will not change its rate of output, firm 1 will behave like a monopolist in determining its profit-maximizing quantity. That is, because q_2 is assumed constant, the marginal revenue equation for firm 1 is

$$MR_1 = \frac{dTR}{dq_1} = \frac{d[(950 - q_1 - q_2) \cdot q_1]}{dq_1} = 950 - q_2 - 2q_1$$

Similarly, firm 2 uses $MR_2 = 950 - q_1 - 2q_2$. In both cases, the profit-maximizing rate of output is determined by setting marginal revenue equal to marginal cost. That is,

$$\text{Firm 1} \qquad 950 - q_2 - 2q_1 = 50$$
$$\text{Firm 2} \qquad 950 - q_1 - 2q_2 = 50$$

Solving each equation for the output of the firm gives

$$\text{Firm 1} \qquad q_1 = 450 - 0.5q_2 \qquad (10\text{-}1)$$
$$\text{Firm 2} \qquad q_2 = 450 - 0.5q_1 \qquad (10\text{-}2)$$

Equations (10-1) and (10-2) show the rate of output for each firm based on the output the managers expect the other firm to produce. For example, if firm 2 is expected to produce 200 units, the profit-maximizing rate of output for firm 1 will be 350 units. Similarly, if firm 1 is expected to produce 200 units, then firm 2 will produce 350 units. Equations (10-1) and (10-2) are referred to as reaction functions because they describe how each firm reacts to the output choice of the other. These reaction functions are portrayed graphically in Figure 10.5.

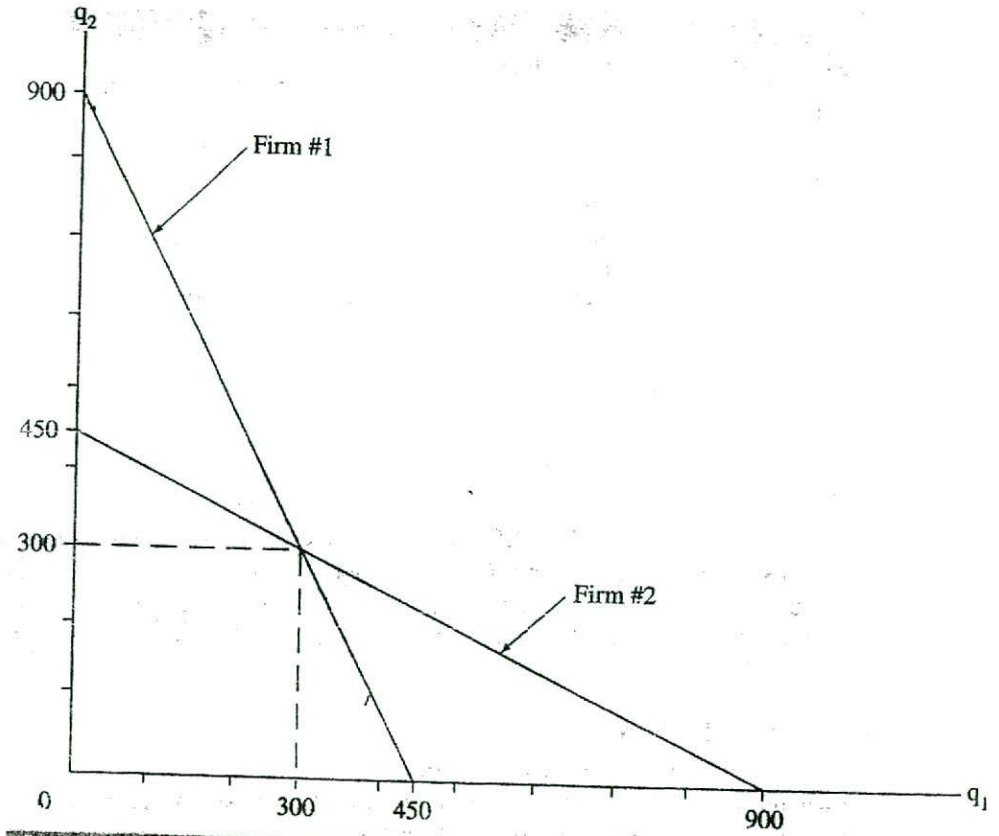
At some point in time, suppose that firm 1 assumes that firm 2 will produce 200 units of output. Based on its reaction function, firm 1 will produce 350 units. If firm 1 produces 350 units, firm 2's reaction function specifies that it should then produce 275 units. But at an output rate of 275 units from firm 2, firm 1's quantity of 350 units is no longer optimal, so it will alter its output based on its reaction function. When firm 1 changes, then firm 2 needs to make an adjustment. As long as the output of one firm is different than that used by the other in selecting its optimal quantity, there will be adjustment.

The market will reach an equilibrium when each firm's output expectation about the other turns out to be correct. Mathematically, this is determined by simultaneously solving equations (10-1) and (10-2). This can be done by substituting $450 - 0.5q_2$ for q_1 in equation (10-2), which gives

$$q_2 = 450 - 0.5(450 - 0.5q_2)$$

or $q_2 = 300$. Substituting this value into equation (10-1) yields $q_1 = 300$. Thus, when each firm produces 300 units, the market has reached an equilibrium. This result can also be shown graphically. Note that in Figure 10.5, the two reaction curves intersect at 300 units of output for each firm. This point of intersection is the graphical equivalent of solving equations (10-1) and (10-2) simultaneously.

The Cournot Model with *n* Firms Although the duopoly market structure is the easiest, the Cournot approach can be used to analyze industries with more than two firms. The mathematics will not be developed here, but for an industry with n firms, the total equilibrium output for a Cournot oligopoly is given by



$$Q_n = Q_c\left(\frac{n}{n+1}\right) \tag{10-3}$$

where $n \geq 1$ and Q_c is the output resulting from a perfectly competitive market.

In the previous section, it was determined that $Q_c = 900$. For the case of a monopoly, $n = 1$. Thus, equation (10-3) implies that a monopolist's output would be 450 units and the combined output of two duopolists ($n = 2$) would be 600 units. These are the same values that were calculated in the previous section. In general, note that as n becomes large, the value of $n/(n + 1)$ approaches unity. This means that as the number of firms in the market increases, the combined output of those firms approaches that of a perfectly competitive market. But as output increases because of more firms participating in the market, the demand equation, $P = 950 - Q_T$, implies that price must decrease. Thus, the Cournot model suggests that increased competition, as measured by the number of firms in the market, drives prices down toward costs.

Table 10.3 shows outputs, prices, and economic profits for different numbers of firms in a Cournot oligopoly. Profits were computed by multiplying output times (price − marginal cost). Note that economic profits decline as the number of firms increases and that maximum profit is obtained when there is a single seller in the market.

Number of Firms	*Total Output*	*Price*	*Profit*
1	450	$500	$202,500
2	600	350	180,000
4	720	230	129,600
8	800	150	80,000
16	847	103	44,891
32	873	77	23,571
64	886	64	12,404
128	893	57	6,251
1,000	899	51	899

Key Concepts

- In the Cournot model, each firm makes its output decision assuming that other firms will produce the same amount as before.
- As the number of firms increases, the Cournot result approaches the equilibrium result for perfect competition.

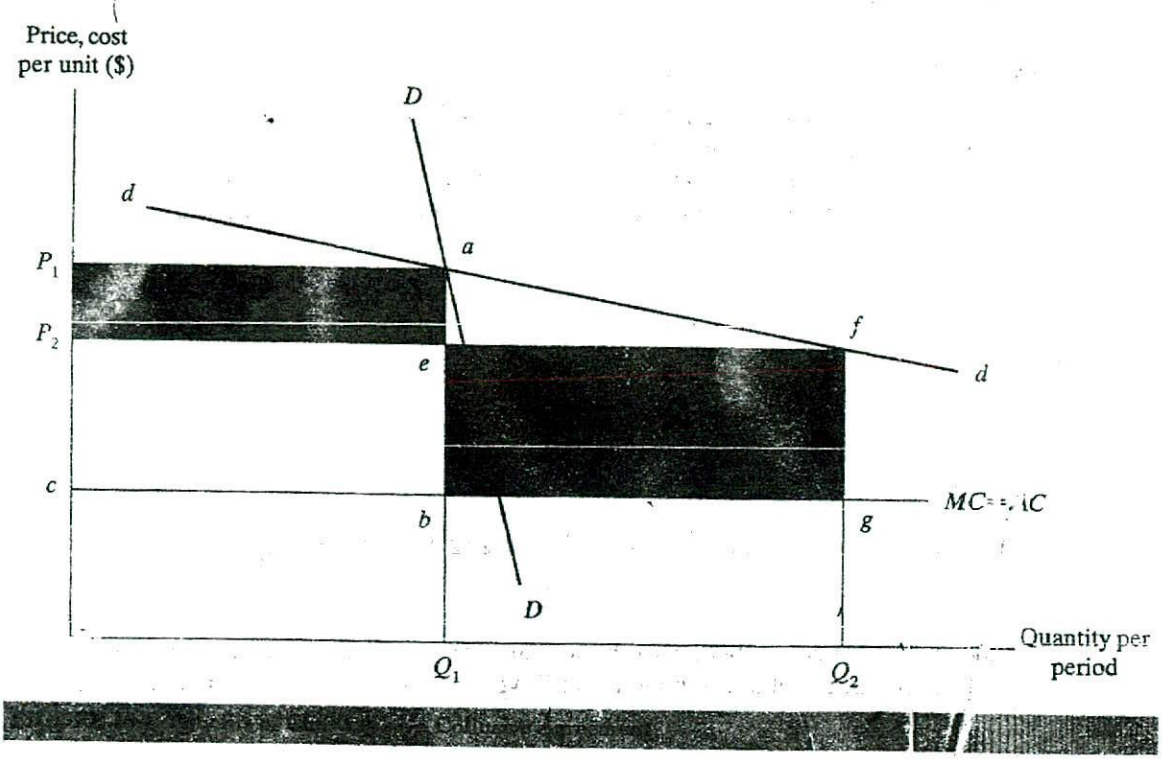
Cartels and Collusion

In oligopolistic industries, vigorous price competition among firms tends to drive prices down and to reduce profits. Consequently, in such industries there is a strong incentive for managers to avoid price competition. One alternative is to collude and set prices at or near the monopoly level.

Although there may be difficulties in formulating an agreement or in dividing the profit, successful collusion can result in substantial benefits for all of the firms involved. Thus, there is a natural tendency for collusion to occur in such industries. This tendency may take the form of explicit price-fixing agreements, price leadership, or other practices that reduce competition between the firms in the market. The exact nature of the collusion in an industry is determined by the specific characteristics of the market and by constraints imposed by government policy. Still, a useful model of oligopoly behavior can be formulated based on the assumption that many managerial decisions are directed toward avoiding active competition and maintaining pricing discipline in the industry.

Collusion and Cheaters Although successful collusion can improve the profitability of all the firms in an industry, any one firm can benefit still more by cheating on the agreement. For example, if the managers in an industry collude and raise prices, an individual firm will be able to increase its share of the market and its total profit by offering a price slightly below that charged by other firms.

The benefits of cheating on a collusive agreement are illustrated by Figure 10.6. Consider a firm participating in an oligopolistic market. Assume that the firms in the market all agree to charge a price P_1. If average and marginal costs are constant and equal, as illustrated by the figure, economic profit earned by each member of the cartel is shown by the rectangle P_1abc.



If every firm in the industry reduces its price, the demand curve faced by the individual firm will be *DD*, as shown in Figure 10.6. This curve is relatively inelastic because any increase in sales for the firm associated with lower prices must result from increased demand for the industry's product.

Now assume that one firm cheats on the collusive agreement and unilaterally reduces its price to P_2. If the price cut is undetected and unmatched, the demand curve faced by this firm will be *dd*, and quantity demanded will increase from Q_1 to Q_2. This increase results from sales of the product to new customers and from sales to customers who previously purchased from other firms.

By selling at a lower price, P_2, the firm will lose profits on the Q_1 units that it was selling at the higher price. But this loss will be more than recouped by profits on the additional sales, $Q_2 - Q_1$. The loss is shown in Figure 10.6 as the shaded rectangle P_1aeP_2. But the gain is the much larger shaded rectangle, *efgb*. Hence, by cheating, the firm is able to earn additional economic profits. A similar opportunity is available to all the other firms in the industry.

As long as the other firms adhere to the price-fixing agreement, the cheater will continue to earn additional profit. Eventually, however, other firms will become aware of the actions of the cheater and will reduce their prices. When this occurs, the price-fixing agreement starts to fall apart. Unless there is a mechanism for restoring price discipline in the industry, the firms may revert to active price competition.

Key Concepts

- By colluding to avoid price competition, firms in oligopolistic markets can increase the total profit to be shared.
- Individual firms can gain even more by setting a slightly lower price and increasing their market share at the expense of other members of the cartel.

Case Study

Raising the Price of Unleavened Bread: The Matzo Conspiracy

For 3,000 years, matzo has been used by Jewish people to commemorate Moses' leading the children of Israel out of bondage in Egypt. This flat, unleavened bread is eaten at Passover to symbolize a departure so hasty that the Israelites couldn't wait for their bread to rise.

For many years, the ancient tradition of eating matzo bread at Passover has provided a small number of producers with a profitable business. Because the product has religious significance, demand for matzo tends to be rather insensitive to price and to changes in economic conditions.

For seven decades, three firms dominated the market for matzo in the United States. But in May 1991, the executive officers of these three firms were indicted by a federal grand jury for colluding to fix the price of matzo. A few months later, Manischewitz, the largest of the three firms, was fined $1,000,000 by a federal judge and also agreed to give several million dollars more in cash and food to charities to settle several private price-fixing cases that were pending.

The suit against Manischewitz alleges that the origin of the collusion was a price war in Chicago involving gefilte fish. When the price cutting on gefilte spread to matzo, executives of the three leading matzo producers met together to arrange a truce that would stabilize prices. Once communication had been established, the executives went a step further and began to meet together on a regular basis to collude on price increases for matzo. Beginning in 1981, they met each fall for five years to decide how much the price of matzo would be raised for the following Passover season. During that period, the price of matzo increased by nearly 38 percent, while the price index for all other food items increased by only 25 percent.

Although the antitrust action ended price fixing of matzo, it did not result in lower prices to consumers. In 1991, Manischewitz acquired one of the two other firms and now controls 90 percent of the U.S. market for matzo. ■

Price Leadership

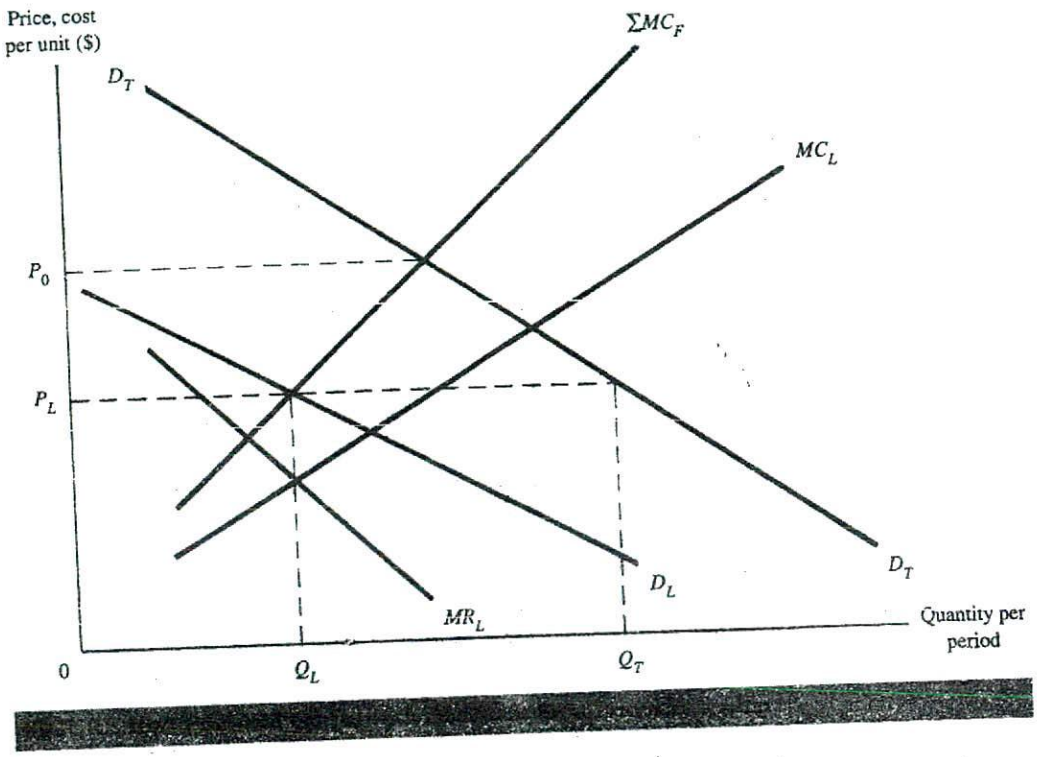
U.S. antitrust laws make explicit collusion difficult and potentially costly. As a result, oligopolists sometimes use other methods to avoid active competition. One of the most prevalent is price leadership. Basically, price leadership occurs when one firm initiates changes in price and the other firms in the industry follow the lead of the first firm. Frequently, the practice of price leadership in an industry first occurred during a period when large price fluctuations and cutthroat pricing were the rule. As a result of price leadership, price changes came to be made infrequently and in ways designed to maintain price discipline in the industry. Two forms of price leadership are considered here: dominant firm and barometric price leadership.

Dominant Firm Price Leadership Consider an industry consisting of a single large firm and several smaller firms. The large firm may have achieved its position by being the first seller in the industry, because it has lower costs resulting from scale economies, or by virtue of superior management skill. Whatever the reason, assume that the firm is now able to dictate prices in the industry. Smaller firms that fail to conform may find themselves in a price war that they cannot survive. However, because the price dictated by the dominant firm is likely to be higher than would result from active competition, the small firms probably will earn more profit by allowing the dominant firm to take the lead in price setting. From the perspective of the industry leader, a closely followed pattern of price leadership eliminates the cost of enforcing industry price discipline. Also, if a large firm is too aggressive in competing in a market, it may be prosecuted for illegal monopolization under antitrust statutes.

Figure 10.7 describes pricing and output decisions with dominant firm price leadership. Let $D_T D_T$ represent total demand. If the small firms in an industry look to the dominant firm to establish price, they can be viewed as price takers. As such, their behavior is similar to firms in perfect competition. If they can sell all they produce at the market price, they maximize profits by producing until price equals marginal cost. The implication of this assumption is that the marginal cost curve for each small firm shows the output that it will produce at various prices established by the dominant firm. Thus, the total output supplied by all the small firms is the sum of their marginal cost curves. This curve is shown in Figure 10.7 as $\Sigma\, MC_F$.

If the dominant firm is content to set a price and let its small rivals supply as much as they want, the large firm can be thought of as supplying the residual demand. For example, if the price is set at P_0, the small firms will meet total market demand and the dominant firm will have no sales because there is no residual demand. It is obvious, therefore, that price will be set below P_0. The demand curve faced by the industry leader can easily be determined for any price level. It is the horizontal distance between the total demand curve, $D_T D_T$, and the $\Sigma\, MC_F$ curve. The leader's demand curve is shown on Figure 10.7 as $D_L D_L$. The associated marginal revenue curve is MR_L.

The curve MC_L is the dominant firm's marginal cost curve. It is shown as being below the $\Sigma\, MC_F$ curve to indicate a cost advantage of the price leader. Because the small firms follow the lead of the dominant seller, the dominant firm acts as a monopolist. Taking $D_L D_L$ as the firm's demand curve, it maximizes profits by choosing the rate of output where marginal revenue equals marginal cost. Thus, output for the price leader will be Q_L. The profit-maximizing price, P_L, is determined by the price leader's demand curve. At this price, output of the small firms is determined from the $D_T D_T$ and $D_L D_L$ curves to be $Q_T - Q_L$ units.



Dominant firm price leadership is less common today than in years past. At one time, firms such as U.S. Steel, Firestone, Alcoa, and IBM were the acknowledged price setters in their markets. Over time, however, market growth, technological change, new U.S. producers, and foreign competition have reduced their dominance. Although still leaders, their relative importance has diminished.

Barometric Price Leadership Price leadership can occur even if there is no dominant firm in a market. Where price changes occur only in response to clear and widely understood changes in market conditions, a pattern of barometric price leadership may evolve. For example, suppose that a union wage settlement has increased labor costs in the industry or that fuel costs have risen. Either of these events will increase costs. One firm in the industry may take the lead in announcing that, due to higher costs, it is necessary to increase prices. Because the price hike reflects industrywide cost changes, other firms are likely to follow suit and increase their prices.

Similarly, if stagnating sales are being experienced by all firms, one seller may announce a price cut to stimulate the demand for its product. If it is well understood that the price reduction is a response to changing market conditions and not an attempt to increase market share at the expense of competitors, the action is unlikely to precipitate a price war. Rather, other firms will match the first firm's price change in an orderly and nonthreatening manner.

With barometric price leadership, it is not necessary that the same firm always function as the price leader. The critical requirement for being a leader is the ability to interpret market conditions and propose price changes that other firms are willing to follow.

Thus, it may be somewhat misleading to define this pattern of behavior as price leadership. Rather, it represents an accepted and legal method of signaling a need for price changes.

Key Concepts

- Dominant firm price leadership involves a single firm setting its profit-maximizing price and the other firms in the industry charging the same price.
- Barometric price leadership is a method of signaling that changes in costs or demand require a price change.

Case Study

Reestablishing Price Discipline in the Steel Industry

Until the early 1960s, U.S. Steel was the leader in setting prices in the steel industry. However, in 1962, a price increase announced by U.S. Steel provoked so much criticism from customers and elected officials, especially President John F. Kennedy, that the firm became less willing to act as the price leader. As a result, the industry evolved from dominant firm to barometric price leadership. This new form involved one firm testing the waters by announcing a price change and then U.S. Steel either confirming or rejecting the change by its reaction.

In 1968, U.S. Steel found that its market share was declining. The company responded by secretly cutting prices to large customers. This action was soon detected by Bethlehem Steel, which cut its posted price of steel from $113.50 to $88.50 per ton. Within three weeks, all of the other major producers, U.S. Steel included, matched Bethlehem's new price.

The lower industry price was not profitable for the industry members. Consequently, U.S. Steel signaled its desire to end the price war by posting a higher price. Bethlehem waited nine days and responded with a slightly lower price than that of U.S. Steel. U.S. Steel quickly dropped its price to Bethlehem's level. Having been given notice that U.S. Steel was once again willing to play by industry rules, Bethlehem announced a price increase to $125 per ton. All of the other major producers quickly followed suit, and industry discipline was restored. Note that the price of $125 per ton was higher than the original price of $113.50. ■

MARKET STRUCTURE AND BARRIERS TO ENTRY

Many factors can contribute to the existence of a particular market structure. However, in the long run, conditions of entry may be the most important determinant. Difficulties encountered in entering an industry are often referred to as *barriers to entry.*

There is some disagreement among economists as to what constitutes a barrier to entry. Bain argues that entry barriers should be defined in terms of any advantage that existing firms hold over potential competitors.[5] In contrast, Stigler contends that for any given rate of output, only those costs that must be borne by new entrants but that are not borne by firms already in the industry should be considered in assessing entry barriers.[6]

Two examples illustrate the difference in philosophy. Suppose that one firm had control over all the iron ore deposits in the United States. As a result, new entrants into the steel industry could get ore only by transporting it from Canada or another foreign supplier. Transportation costs would cause potential competitors to have higher costs of producing steel than those of the existing firm. This disadvantage could prevent the new firms from successfully entering the market. Both the Bain and the Stigler criteria for a barrier to entry are satisfied in this example. Alternatively, assume that iron ore deposits are equally available to the existing and the potential new firms, but that the existing firm is large enough to take advantage of highly efficient production technologies. If new entrants build plants of small scale, their costs may be so high that they cannot sell steel at a price competitive with the established firm. That is, successful entry requires construction of plants that are large enough to take advantage of economies of scale. Bain would consider this condition a barrier to entry because of the difficulties in coordinating and raising capital for large-scale entry. However, Stigler's definition would not recognize scale economies as an entry barrier because the old and the new firms both face the same cost conditions. That is, for any given rate of output produced, the cost per unit would be the same for the new entrant as for an existing firm.

From a strictly conceptual point of view, the Stigler position that entry barriers should be confined to problems faced by new, but not existing, firms has appeal. But the Bain definition is the more useful of the two approaches. By including all factors that impede entry into an industry, it provides a better framework for understanding the determinants of market structure.

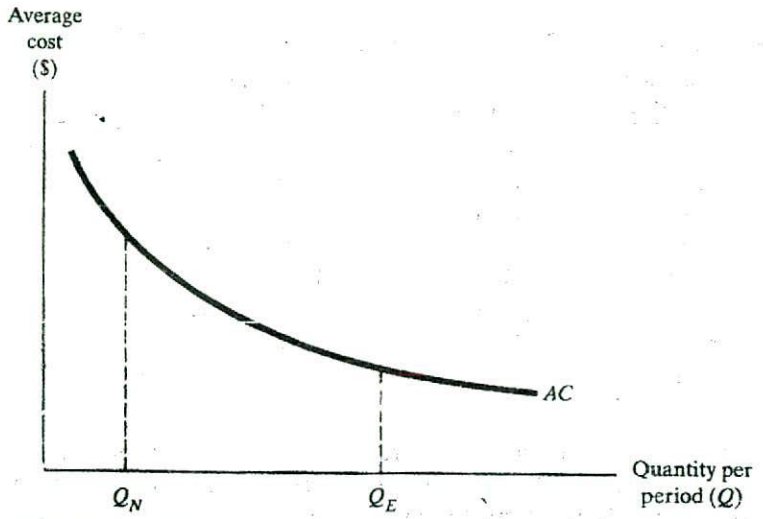
Sources of Barriers to Entry

Although there are many possible factors that restrict entry, this discussion focuses on four of the most important. The first is product differentiation. A firm that has convinced consumers that its product is significantly better than the product of new entrants has an advantage. The new firm may be forced to sell its product at a lower price that may not generate an adequate profit. There is no requirement that the product really be superior, only that consumers perceive it as more desirable. For example, some aspirin manufacturers have extolled the virtues of their product for decades. The result is that the typical consumer has a subjective feeling that Bayer and other well-advertised aspirins are superior to other brands—despite their being chemically identical. The ability of firms such as Bayer to differentiate their product allows them to capture a larger share of the market, charge higher prices, and earn economic profit. Even if new firms were to spend as much on advertising each year as Bayer, it would be unlikely that they could ever catch up. Bayer's long tenure in the market has generated strong consumer loyalties that are difficult for new firms to overcome. Thus, a barrier to entry exists.

---

[5] J. S. Bain, *Barriers to New Competition* (Cambridge, MA.: Harvard University Press, 1956), pp. 3–5.

[6] G J. Stigler, *The Organization of Industry* (Burr Ridge, IL.: Richard D. Irwin, 1968), pp. 67–70.



A second restriction on entry is control of inputs by existing suppliers. If potential entrants cannot easily obtain the capital, raw materials, and labor needed to produce their product, entry may be difficult. Examples include scarcity of natural resources, locational advantages, and managerial talent.

Legal restrictions such as patent protection are a third source of entry barriers. In some industries, patents held by existing firms make it virtually impossible for other businesses to produce a comparable product. Exclusive franchises granted by the government are another form of legal restriction. For example, an electric company may have a legal right to be the sole supplier of electricity in its service area.

Scale economies are a fourth source of barriers to entry. As shown in Figure 10.8, if the production process exhibits economies of scale, a large, existing firm, producing at an output rate of Q_E, will have lower average costs than a new firm attempting to enter the industry on a small scale, such as Q_N. As a result, the new entrant may not be able to operate profitably at a price that allows existing firms to earn substantial economic profit. At the same time, any effort to enter the industry on a larger scale may be frustrated by difficulties of obtaining capital and putting together the necessary organization. Thus, the ability of existing firms to expand gradually as compared to the need for new entrants to start out with considerable production capacity can be a substantial advantage for existing firms. A good example is the automobile industry. Because of the economies of scale associated with the manufacture and sale of automobiles, no new domestic producer has successfully entered the industry in over sixty years.

Spectrum of Market Structures

As discussed in chapter 9, there are few markets that can properly be categorized as being either perfectly competitive or true monopolies. The majority of markets lies in the intermediate range characterized by monopolistic competition and oligopoly.

TABLE 10.4 Market Structures of Selected Industries

Perfect Competition	*Monopolistic Competition*	*Oligopoly*	*Dominant Firm Oligopoly*	*Monopoly*
Agriculture	Movie theaters	Automobiles	Aircraft (Boeing)	Electric and gas utilities
Futures trading	Printing	Cereal	Canned soup (Campbell)	
	Retailing	Cigarettes	Film (Kodak)	
	Restaurants	Beer	Detergents (Procter & Gamble	
	Dresses	Oil refining		
	Poultry	Steel		
	Yarns	Newspapers		
	Sheet metal	Chewing gum		

Table 10.4 shows a spectrum of market structures and representative industries in each category. Although agriculture is listed under perfect competition, it does not meet all the requirements for that designation. Rather, it is an industry that comes very close to being perfectly competitive. Similarly, electric and gas companies may have legal monopolies in their service areas, but they still face some competition from substitute services. Thus, even they are not true monopolies.

Note that two types of oligopolies are shown in Table 10.4. The first type (listed in the third column) consists of a small number of firms of nearly equal size. The second type (shown in the fourth column) has a dominant firm. These dominant firms are shown in parentheses with each industry and supply a large share of total industry output. For example, several firms manufacture film in the United States. However, Kodak is still the dominant firm in that industry. The market shares of other U.S. film manufacturers are small.

Key Concepts

- In the long run, barriers to entry may be the most important determinant of market structure.
- Sources of entry barriers include control of scarce inputs, product differentiation, legal factors, and scale economies.
- Most markets resemble monopolistic competition and oligopoly more than perfect competition and monopoly.

ase Study

rriers to Entry: New York Taxi Drivers and Australian Fishermen

To limit the number and regulate the operations of taxicab drivers, in 1937 the city of New York required all cab drivers to purchase a license. These licenses were called medallions and were sold for $10 each. At that time, 11,787 medallions were issued and no additional licenses have been granted in the intervening years. Thus, those who own

the medallions are part of a taxi service monopoly in New York. A legal barrier to entry exists, which prevents others from entering this market.

Although the supply of licensed taxicabs has not changed, the demand for taxi service has increased dramatically since 1937. As a result, the medallions have become extremely valuable, because ownership of a medallion confers upon the owner an opportunity to earn economic profits. Essentially, a taxicab license is worth the present value of the future profit stream that it can generate. By the 1990s, New York City taxi medallions were selling for over $200,000.

The legal barrier to entry in New York City has resulted in an excess demand for taxi service that has caused problems. Nonlicensed cabs cruise the streets of New York and illegally pick up passengers. Sometimes their presence results in violent confrontations with medallion drivers. Another difficulty is that licensed drivers, with ample opportunities in downtown New York, are often unwilling to pick up or deliver customers to bad neighborhoods or outlying areas. One solution has been to authorize a limited number of nonmedallion cars. These vehicles can respond to a telephone request for taxi service, but are not allowed to pick up people who want to hail a cab from the street.

New York taxis are not the only example of monopoly resulting from a barrier to entry created by governmental restrictions. In Australia, offshore fishing is regulated by the Department of Fisheries. In order to fish Australia's coastal waters, fishermen must obtain a license from the department. But only a limited number of licenses are issued and, once obtained, the license grants a permanent right to fish in a specified area. The purpose of licensing is to prevent the region from being depleted by overfishing.

Because their number is restricted, these licenses can be extremely valuable. For example, one fishing area off the southern coast of Australia is a prolific source of giant prawns. But only 39 boats are authorized to operate there. Consequently, licenses have sold for more than $900,000. This high price can be justified by the $400,000 to $500,000 in prawns that can be harvested during each year's 120-day fishing season. ■

ADVERTISING

Of the market structures listed in Table 10.4, advertising is most important to firms in monopolistically competitive and oligopolistic markets. By definition, firms in perfect competition sell an undifferentiated product. Because they can sell all they produce at the market price, they have no need to advertise. At the other extreme, as a single seller, a monopolist does not need to advertise to maintain its market share. Rather, dollars spent on advertising by monopolists and near monopolists are intended to increase total demand for the firm's product. Usually, advertising expenditures are a relatively small fraction of total revenues.

Monopolistic competition differs from perfect competition in that monopolistically competitive firms sell a slightly differentiated product. Advertising is one way of achieving product differentiation. Advertising also is important in oligopolistic markets where it is intended to increase a firm's market share. Competition through advertising may also serve as a partial substitute for active price competition in such industries.

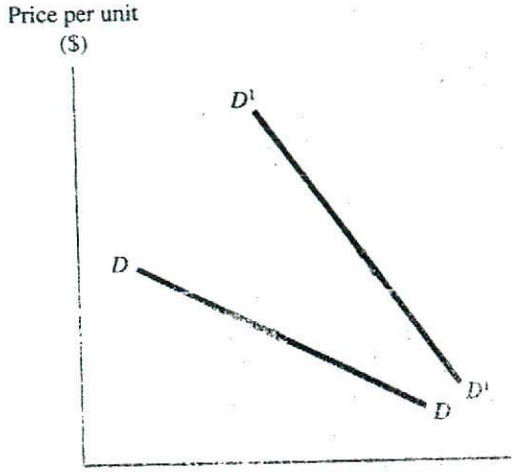


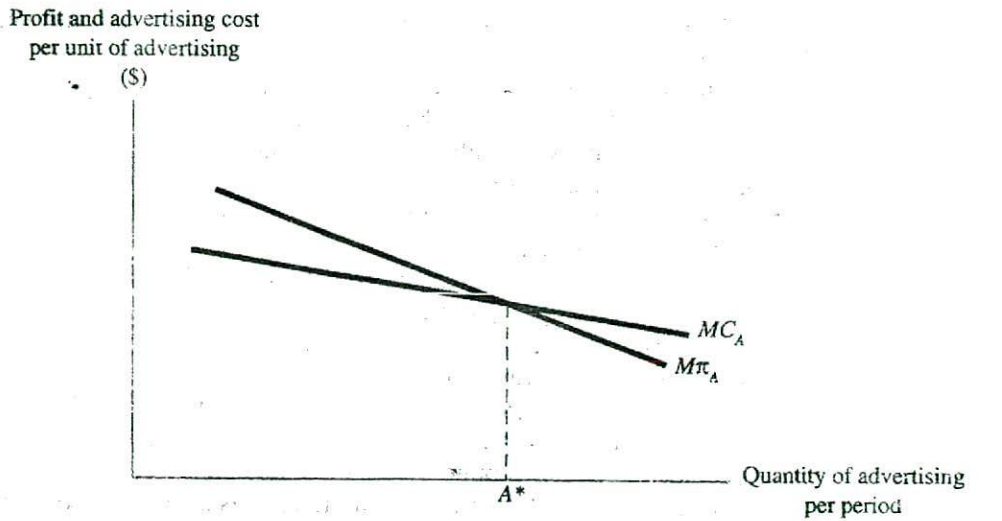
FIGURE 10.9 The Effect of Advertising

Consequently, firms in oligopolistic markets often spend a large proportion of their total revenues on advertising.

How should managers determine the amount to spend on advertising? In Figure 10.9, let DD be the demand curve before advertising and D^1D^1 the curve after advertising. In analytical terms, the objective of advertising is to shift the demand curve to the right and make demand less elastic. Note that the figure implies that if advertising is effective, more of the product can be sold at each price, and price changes have a smaller effect on quantity demanded.

But as with inputs in the production process, the marginal effectiveness of advertising diminishes as additional advertising dollars are spent. That is, advertising is subject to diminishing marginal returns. There may be several reasons why this occurs. First, to the extent that advertising conveys information to consumers, there may be a limit to the amount of information a person wants to obtain or is able to process. If advertising proceeds beyond this limit, little additional benefit will be received. Second, consumers may become irritated if a product is advertised too often. If this happens, they may react by ignoring the information or by developing a distaste for the product. Finally, if one firm in an oligopoly increases its advertising rate, it may cause other firms to respond with extensive advertising campaigns of their own. Thus, the advertising may be primarily self-canceling and no single firm will benefit.

If advertising can increase demand but at the same time is subject to diminishing marginal returns, there is an optimal rate of advertising expenditure for the firm, as shown in Figure 10.10. The line $M\pi_A$ represents the marginal profitability of advertising. That is, it shows the extra dollars in profit earned on sales resulting from an additional unit of advertising (e.g., a 1-minute television commercial or a page in a magazine). Note that the line is downward sloping because of diminishing marginal returns to advertising. The line MC_A depicts the marginal cost of a unit of advertising. It is also shown as downward sloping because quantity discounts are frequently available to large advertisers. For example, Procter & Gamble spends over $1 billion on advertising each year. The firm's volume of advertising enables it to purchase time and space at a lower cost than is available to smaller firms.



The advertising decision is made like other resource allocation decisions—the activity is increased until marginal benefit equals marginal cost. In the example, this occurs at A^*, where the extra profit resulting from an additional unit of advertising just equals the cost of buying one more unit. Beyond that point, additional advertising expenditures would exceed the additional profit they generate for the firm.

SUMMARY

Chamberlin's model of monopolistic competition assumes ease of entry and exit and a large number of small sellers. It differs from perfect competition by viewing sellers as providing products that are slightly differentiated. Thus, firms have some control over price. In the short run, profits are maximized by equating marginal revenue and marginal cost, and there are economic profits. In the long run, entry of new firms causes prices to fall and eliminates economic profits. The theory of monopolistic competition has a limited scope of application, but the model makes a useful contribution to economic analysis by calling attention to the importance of product differentiation.

Oligopolistic market structures have many buyers but only a small number of sellers. The product may be either differentiated or undifferentiated. Typically, entry into the industry is somewhat difficult. An important difference between oligopoly and other market structures is that oligopolists recognize their interdependence. Thus, in making decisions, managers must consider the effect on other firms and the probable response of those firms.

There is no single theory that describes all aspects of oligopoly behavior. However, specific models capture certain elements. The kinked demand curve model assumes that competitors will follow price reductions but not price increases. The implication is that price changes will be infrequent in oligopolistic markets.

In the Cournot model of oligopoly, each firm makes its profit-maximizing output decision on the assumption that other firms will not change their rate of output. Using the Cournot assumption, it can be shown that output increases and the price decreases as the number of firms in the market increases.

Oligopolists have an incentive to collude and avoid aggressive competition. By collusion, the firms can increase their total profit. But individual firms can gain even more if they cheat on a collusive agreement. The success of a cartel depends on its ability to detect and punish cheaters.

Price leadership is a substitute for illegal collusion. Where there is a dominant firm, that supplier will charge its profit-maximizing price and smaller firms will charge the same price. Barometric price leadership involves firms signaling each other that changes in demand or costs require a price change.

Barriers to entry are probably the most important long-run determinant of market structure. Entry barriers can result from control of scarce inputs, product differentiation, legal factors, and economies of scale.

To maximize profit, the firm should increase its advertising expenditures until the marginal profit generated by advertising equals the cost of the last unit of advertising.

Discussion Questions

10-1. Is product differentiation important in the breakfast cereal industry? Explain.

10-2. Which of the following markets could be considered monopolistically competitive? Explain.
   a. Network television.
   b. Low-priced pens.
   c. Restaurants.
   d. Automobiles.

10-3. In monopolistically competitive markets, why is the firm's demand curve assumed to be relatively more elastic in the long run than in the short run?

10-4. Monopolistic competition assumes slightly differentiated products, but the average and marginal cost curves used in the analysis are those of a typical or representative firm. Are these assumptions consistent? Explain.

10-5. Suppose that there are 5,000 firms selling candy by mail to people in the Pacific Northwest. Could this market be characterized as monopolistically competitive? If 5,000 firms are selling concrete in the same region, could that market be considered monopolistically competitive? Explain.

10-6. Why might oligopolists be more likely to match a price cut than a price increase by a competitor?

10-7. Contracts for electric generating equipment are often awarded on the basis of sealed bids. As an employee of the U.S. Department of Justice, what would you look for as an indication of price fixing in the electric machinery industry?

10-8. Why would smaller firms be content to let a large firm practice dominant firm price leadership in an industry?

10-9. Accumulated experience may allow firms that have been producing for many years to have lower costs than new entrants in a market. Would this learning by doing be a barrier to entry as defined by Bain? What about using Stigler's definition of entry barriers?

10-10. Are service industries more likely to be near the monopoly or the competitive end of the spectrum of market structures? Why?

Problems

10-1. In Gotham City the movie market is monopolistically competitive. In the long run, the demand for movies at the Silver Screen theater is given by the equation

$$P = 5.00 - 0.002Q$$

where Q is the number of paid admissions per month. The average cost function is given by

$$AC = 6.00 - 0.004Q + 0.000001Q^2$$

a. To maximize profit, what price should the managers of Silver Screen charge? What will be the number of paid admissions per month?

b. How much economic profit will the firm earn?

10-2. The demand equation for a firm operating in a monopolistically competitive market is given by $P = 4.75 - 0.2Q$. Average cost for the firm is given by $AC = 5 - 0.3Q + 0.01Q^2$. The firm is in long-run equilibrium.

a. What is the profit-maximizing price and quantity?

b. How much economic profit will the firm earn?

10-3. EnviroEast can produce recycled paper at a constant marginal cost of \$0.50 per pound. Currently, the firm is selling 500,000 pounds each year at \$0.60 per pound. Managers of EnviroEast are considering increasing the price to \$0.90 per pound. Demand elasticity is constant and equals -0.6 if the price increase is matched by competitors and -4.0 if it is not matched. Management believes there is a 70 percent chance that other firms will follow EnviroEast's lead and increase their prices.

a. If managers are risk neutral, should the proposed price change be implemented? Explain.

b. Write an equation for the expected change in profit as a function of the probability that the price increase will be matched by the other firms. What probability would make a risk-neutral manager indifferent to the change?

10-4. The price of steel is currently at \$400 per ton. Pennsylvania Steel faces a kinked demand curve with a demand equation

$$P = 600 - 0.5Q$$

for prices above the present price of \$400 and

$$P = 700 - 0.75Q$$

for prices below \$400. The firm's marginal cost curve is given by an equation with the general form

$$MC = a + bQ$$

a. If $a = 50$ and $b = 0.25$, graph the demand, marginal revenue, and marginal cost curves. Using the graph, determine the profit-maximizing quantity for Pennsylvania Steel.

b. Starting from $a = 50$, how much can the constant term of the marginal cost equation increase before the profit-maximizing quantity decreases? How much can the coefficient decrease before the profit-maximizing quantity increases?

10-5. If other firms in an oligopolistic industry do not respond to changes in the price of a firm's product, the demand curve is $Q = 700 - 50P$. However, if other firms always match the firm's price, the demand curve is $Q' = 200 - 10P$.

a. If the firm's marginal cost is \$8.00, what is the profit-maximizing quantity and price?

b. If the firm's marginal cost increases to \$11.50, what will be the profit-maximizing quantity and price?

10-6. At present, the price of copper tubing is \$1 per foot. Lyon, Inc. is considering entering the industry by building a facility that will produce 40 million feet of tubing per year. It is estimated that the average cost curve for manufacturing copper tubing is given by the equation

$$AC = 1.21 - 0.010Q + 0.0001Q^2$$

where AC is average cost per foot (including a normal profit) and Q is millions of feet per year.

a. If the price of copper tubing remains unchanged, should Lyon, Inc. build the planned production facility? Why or why not?

b. At a price of \$1 per foot, what is the rate of output per year necessary to earn at least a normal profit?

c. Suppose the probability is 0.7 that the price will stay at \$1 per pound and 0.3 that Lyon's entry will cause the price to drop to \$0.90 per pound. If managers are risk neutral, should they build the production facility? Why or why not?

10-7. The demand for spring water is given by $P = 1000 - Q_T$, where Q_T is the total amount of the product sold. The marginal cost is zero. The firms in the market behave like Cournot oligopolists.

a. If there are two firms, determine the reaction functions for each firm. What will be the equilibrium price and equilibrium total output?

b. What would be the price and output for a monopolist?

c. What would be the price and output if there were six firms in the market?

10-8. Two firms face a demand equation given by $P = 200{,}000 - 6(q_1 + q_2)$, where q_1 and q_2 are the outputs of the two firms. The total cost equations for the two firms are given by

$$TC_1 = 8{,}000\,q_1 \text{ and } TC_2 = 8{,}000\,q_2$$

a. If each of the firms sets its own output rate to maximize its profits, assuming that the other firm holds its rate of output constant, what will be the equilibrium price?

b. How much output will each firm produce?

c. How much profit will each firm earn?

d. If the firms collude, what will be the monopoly price and output?

e. If profits from collusion are shared equally, how much profit will each firm earn?

10-9. Wyngate, a small manufacturing firm, is considering building a plant capable of producing 25 million wood pencils per year. Economies of scale are the only im-

portant barrier to entry in the industry. It is estimated that the average cost function is given by

$$AC = 100{,}000 - 1{,}000Q + 10.0Q^2$$

where Q is measured in millions of pencils per year. The current wholesale price of pencils is \$75,000 per million, and this price will not be affected by the entry of Wyngate into the industry. What is the minimum output rate necessary to allow a firm to earn at least a normal profit? Will Wyngate be able to successfully enter the industry?

Problems Requiring Calculus

10-10. Southern, Inc. operates in a monopolistically competitive market. The demand equation faced by Southern is given by $P = 350 - Q$, and the firm's long-run total cost equation is given by $TC = 355Q - 2Q^2 + 0.05Q^3$.
   a. What are the equilibrium price and rate of output for the firm?
   b. Compute the economic profit earned by the firm.
   c. Determine whether the equilibrium rate of output calculated in part (a) meets the marginal revenue equals marginal cost test.

10-11. The demand for spring water is $P = 1{,}000 - Q_T$, and marginal cost is zero. There are two firms in the market, and each firm believes that the other will respond to a one-unit decrease in output by increasing its output by one-half unit. For example, if firm 1 reduces output by one unit, firm 2 will respond by increasing its output by one-half unit.
   a. Compute the marginal revenue equations for each firm.
   b. Compute the reaction functions for each firm.
   c. What will be the equilibrium price and the total output for the industry? How do the price and output compare to those computed for the Cournot duopoly in problem 10-7?

10-12. For Jensen Associates, profit as a function of advertising is given by the following equation:

$$\text{Total Profit} = 500 + 50A - A^2$$

where A is units of advertising and each advertising unit costs \$4.
   a. Mathematically determine the optimal rate of advertising.
   b. Now suppose that quantity discounts are available for purchases of advertising and that the marginal cost of an additional unit of advertising is given by

$$MC_A = 60 - 3A$$

Determine the profit maximizing rate of advertising. Explain.