CHAPTER 7

Production

How is Boeing's output of airplanes related to the use of various inputs?

Chapter Outline

Learning Objectives

- Establish the relationship between inputs and output.
- Distinguish between variable and fixed inputs.
- Define total, average, and marginal product.
- Understand the Law of Diminishing Marginal Returns.
- Investigate the ability of a firm to vary its output in the long run when all inputs are variable.
- Explore returns to scale: how a firm's output response is affected by a proportionate change in all inputs.
- Overview how production relationships can be estimated through surveys, experimentation, or regression analysis.

In Chapters 3 through 6 we concentrated on consumer behavior, with the supply of goods taken for granted. Now we begin to analyze the factors that determine the quantities of goods firms will produce and offer for sale.

We begin our examination of the supply side of the market by assuming that firms maximize profit. If a firm is interested in maximizing profit, two important steps must be taken. First, for any potential output level, total cost needs to be minimized. In other words, no more resources than necessary should be employed to produce any given level of output. Second, having minimized the cost of producing a given output, the firm must select the price and corresponding output level that maximizes profit. As we will see, the optimal price–output choice will depend on the market structure in which the firm operates.

This chapter and Chapter 8 focus on the first step that a firm must take to maximize profit. Namely, we examine the firm's technology and the input prices the firm confronts to develop an intuitive rule for minimizing the total cost of producing a given output level. Chapters 9 through 15 address the second step toward profit maximization—choosing the right price–output combination once the total cost of producing a given output level has been minimized.

To arrive at an intuitive rule for cost minimization, the logical starting place is to identify the underlying technological relationships between inputs employed and output produced. This chapter explains how economists represent the technological possibilities available to the firm. The productivity of inputs is an important determinant of output: it specifies how much can be produced. As we will see in later chapters, the productivity of inputs underlies both the cost curves of the firm and the firm's demand curves for inputs.

7.1 RELATING OUTPUT TO INPUTS

FACTORS OF PRODUCTION
inputs or ingredients mixed together by a firm through its technology to produce output

Inputs—sometimes called **factors of production**—are the ingredients mixed together by a firm through its technology to produce output. For example, a motion picture studio uses inputs such as producers, directors, actors, costume and sound designers, technicians, and the capital invested in its lots, sound stages, and equipment to produce movies.

Inputs may be defined broadly or narrowly. A broad definition might categorize all inputs as either labor, land, raw materials, or capital. When considering some questions, however, it may be helpful to use more narrow subdivisions within the broader categories. For example, the labor inputs employed by a firm might include engineers, accountants, programmers, secretaries, and managers. Raw materials may involve electricity, fuel, and water. Capital inputs may include buildings, trucks, robots, and automated assembly lines.

For any good, the existing technology ultimately determines the maximum amount of output a firm can produce with specified quantities of inputs. By *existing technology*, we mean the technical or organizational "recipes" regarding the various ways a product can be produced. The **production function** summarizes the characteristics of existing technology. The production function is a relationship between inputs and output: it identifies the maximum output that can be produced per time period by each specific combination of inputs.

PRODUCTION FUNCTION
a relationship between inputs and output that identifies the maximum output that can be produced per time period by each specific combination of inputs

Consider the case of a firm that employs two inputs, labor (L) and capital (K), to produce output (Q). Input usage and output are measured as *flows*: for example, the units of capital and labor employed *per day* and the firm's *daily* output. For simplicity, however, we generally will omit the time period and refer to units of inputs or output rather than units of inputs or output per relevant time period.

Mathematically, the firm's production function can be written as

$$Q = f(L,K).$$

TECHNOLOGICALLY EFFICIENT
a condition in which the firm produces the maximum output from any given combination of labor and capital inputs

This function indicates what is **technologically efficient**—the maximum output the firm can produce from any given combination of labor and capital inputs. The production function identifies the physical constraints with which the firm must deal. We assume that the firm knows the production function for the good it produces and always uses this knowledge to achieve maximum output from whatever combination of inputs it employs. This assumption of technological efficiency may not always be valid, but there is reason for believing it to be generally correct. Any firm operating in a technologically inefficient way is not making as much money as possible. The firm's cost of using a given level of inputs is the same whether or not it uses the inputs wisely, but the revenue from the sale of the product (and hence the profit) will be greatest when the firm produces the maximum output given these

inputs. Consequently, any profit-oriented firm has an incentive to seek out and use the best available production technique.

7.2 PRODUCTION WHEN ONLY ONE INPUT IS VARIABLE[1]

Naturally, the example of a firm using the two inputs of labor and capital to produce output is exceedingly simple and glosses over many of the subtleties of real-world production technologies. Still, this simple example allows us to illustrate several key features of the relationship between inputs and output that does characterize real-world production. One of these features is what happens to output when a firm can vary the use of only one of its inputs over a given time period.

FIXED INPUTS
resources a firm cannot feasibly vary over the time period involved

Resources that a firm cannot feasibly vary over the time period involved are referred to as **fixed inputs**. These inputs need not be fixed in the sense that varying their use is literally impossible; rather, they are any inputs that are prohibitively costly to alter in a short time period. For example, a commercial real estate developer in New York City may be largely unable to supply additional office space over the coming month in response to an increase in market demand. This is because acquiring land and/or building permits in New York over such a short time frame is virtually impossible. DaimlerChrysler's physical plant provides another example. In response to strong consumer demand, DaimlerChrysler might be able to expand capacity for production of its PT Cruiser car in a month. Doing so, however, would require around-the-clock employment of large numbers of engineers and contractors at exorbitant cost. In that case, practically speaking, the physical manufacturing plant associated with the PT Cruiser car is a fixed input that Chrysler will not vary in the event that quick output adjustments are required.

TOTAL PRODUCT
the total output of the firm

Suppose that in our simple scenario, the firm is stuck with a certain amount of capital for the time being and can vary only the number of workers—the amount of labor—that it employs. For the sake of simplicity, we assume that capital is held constant at 3 units and examine how output or **total product** varies as the firm employs different quantities of labor.

Table 7.1 shows a hypothetical relationship between output and various labor quantities. The first column is included merely to emphasize that the amount of capital input is held

TABLE 7.1 PRODUCTION WITH ONE VARIABLE INPUT

Amount of Capital	Amount of Labor	Total Product	Average Product of Labor	Marginal Product of Labor
3	0	0	—	—
3	1	5	5	5
3	2	18	9	13
3	3	30	10	12
3	4	40	10	10
3	5	45	9	5
3	6	48	8	3
3	7	49	7	1
3	8	49	6.1	0
3	9	45	5	−4

[1]A mathematical treatment of some of the material in this section is given in the appendix at the back of the book (pages xxx–xxx).

constant at 3 units regardless of the labor used. The second and third columns contain the important data, showing how much total product can be produced with alternative quantities of labor. With zero workers, total product is zero. As the amount of labor increases, total product rises. One worker combined with 3 units of capital results in a total product of 5, using 2 workers raises output to 18, 3 workers further increases output to 30, and so on. There is, however, a limit to the total product that the firm can produce by increasing labor input if capital input is held constant at 3 units. In our example the limit is reached when 8 workers are employed and total product is 49. The eighth worker adds nothing to output, and using 9 workers actually causes output to fall.

Although these figures are hypothetical, the general relationship they illustrate is quite common. To examine the relationship further, we introduce the concepts of average product and marginal product of an input. The **average product** of an input is defined as the total output (or total product) divided by the amount of the input used to produce that output. For example, 3 workers produce 30 units of total product, so the average product of labor is 10 units of output at that employment level. The average product for each quantity of labor is therefore derived by dividing the total product in column 3 by the corresponding amount of labor in column 2. Note that total product, and thus the average product of labor, depends on the amount of other inputs—in this case, capital—being used and that the amount of nonlabor inputs is held constant throughout Table 7.1.

AVERAGE PRODUCT the total output (or total product) divided by the amount of the input used to produce that output

The **marginal product** of an input represents the change in total output resulting from a one-unit change in the amount of the input, holding the quantities of other inputs constant. To illustrate, when labor is increased from 4 to 5 units, total output rises from 40 to 45, or by 5 units. So the marginal product of labor, when the fifth worker is employed, is 5 units of output. What the marginal product of an input measures should be thoroughly understood. In many applications it is the crucial economic variable, because most production decisions relate to whether a little more or a little less of an input should be employed. The way total output responds to this variation is what the marginal product measures.[2]

MARGINAL PRODUCT the change in total output that results from a one-unit change in the amount of an input, holding the quantities of other inputs constant

Total, Average, and Marginal Product Curves

The information from Table 7.1 can be conveniently graphed. Figure 7.1 shows the result. (We have assumed that labor and output are divisible into smaller units in drawing the graphs, so the relationships are smooth curves rather than 10 discrete points.) The total product (TP) curve in Figure 7.1a shows how the output varies with the quantity of labor employed. Just as indicated in Table 7.1, output increases as more labor is used and reaches a maximum at 49 units, when 8 workers are employed; beyond 8 workers, output declines.

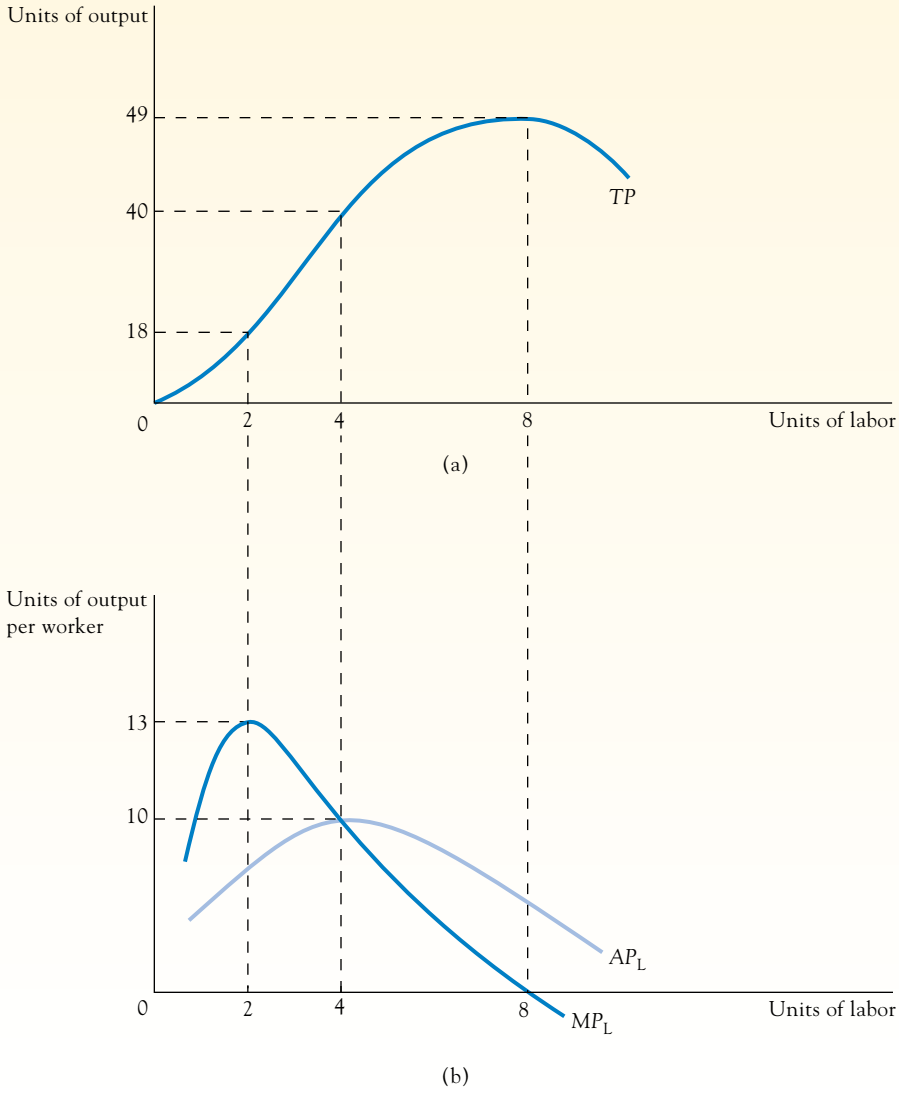
Figure 7.1b shows the average product (AP_L) and marginal product (MP_L) curves for labor. Note that these curves measure the output per unit of input on the vertical axis rather than total product, which is what the vertical axis measures in Figure 7.1a. As employment of labor increases, MP_L increases at first, reaches a maximum at 2 workers, and then declines. The average product of labor also increases at low levels of employment, reaches a maximum height of 10 units per worker, and then declines.

The two panels of Figure 7.1 highlight the relationship between total and marginal product. As long as marginal product is positive, total product rises. That is, as long as an extra unit of labor produces some extra output (however small), the total amount produced increases (up through 7 units of labor in Figure 7.1). When marginal product is negative, total product falls (beyond 8 units of labor), and when marginal product is zero, total product is at its maximum (at 8 units of labor).

[2]Note that the marginal product figures in Table 7.1 pertain to the interval between the indicated amount of labor and 1 unit less. Thus, the marginal product at 4 units of labor is 10, because total output rises from 30 to 40 when labor increases from 3 to 4 units.

FIGURE 7.1

Total, Average, and Marginal Product Curves
(a) The total product curve shows the output produced with various amounts of labor, assuming that other inputs are held constant. (b) Average and marginal product curves are derived from the total product curve.



A rational producer, of course, will never operate where marginal product is negative (beyond 8 workers in Figure 7.1). This is so because employing a variable input at a level where its marginal product is negative is technologically inefficient. The firm can increase its total product and lower its production cost (provided that the price of the variable input is positive) by using less of the variable input.

The Relationship Between Average and Marginal Product Curves

A definite relationship exists between the average and marginal product curves. When marginal product is greater than average product, average product must be increasing, as is shown between 1 and 4 units of labor in Figure 7.1b. This relationship follows directly from

the meaning of the terms. If the addition to total product (marginal product) is greater than the average, the average must rise. Think of the average height of people in a room. If another person enters who is taller than the average (the marginal height of the extra person is greater than the average), the average height will increase. Similarly, when marginal product is less than average product, average product must decrease, as is shown for labor beyond 4 units in the diagram. Because marginal product is greater than average product when the average is rising, and less than average product when the average is falling, marginal and average products will be equal when average product is at a maximum.

APPLICATION 7.1 MARGINAL AND AVERAGE PRODUCTS IN MAJOR LEAGUE BASEBALL

During the 2001 Major League Baseball season, Barry Bonds hit 73 home runs while playing for the San Francisco Giants. In 2002, Bonds hit 46 home runs—less than his own total the previous season but more than the league-leading total of a hitter in most other preceding years.

In terms of home runs, Bond's marginal (home runs per season) product declined from 73 to 46 between the 2001 and 2002 seasons. However, Bond's average (home runs per season) product increased from 35 in 2001 to 36 in 2002—the average is based on Bond's performance since he entered Major League Baseball in 1986.

How could Barry Bonds' marginal (home runs per season) product decline between 2001 and 2002 while his average (home runs per season) product increase? The reason is straightforward. Since Bond's marginal product in 2002 (46 home runs) exceeded the average product (35 home runs) through his preceding seasons in Major League Baseball, Bond's average product ended up being pulled from 35 to 36 by the 46-home-run season.

The Geometry of Product Curves

As we saw in discussing Table 7.1, knowing how total output varies with the quantity of the variable input allows us to derive the average and marginal product relationships. Similarly, we can use geometrical relationships to derive the average and marginal product curves from the total product curve.

Figure 7.2 illustrates how average product is derived geometrically from the total product curve, *TP*. The average product of labor is total output divided by the total quantity of labor. At point *B* on the total product curve, average product is equal to Q_2/L_2, or 9 units of output per worker. Note, however, that Q_2/L_2 equals the slope of the straight-line segment 0*B* drawn from the origin to point *B* on the total product curve. *Thus, the average product at a particular point is shown geometrically by the slope of a straight line from the origin to that point on the total product curve.*

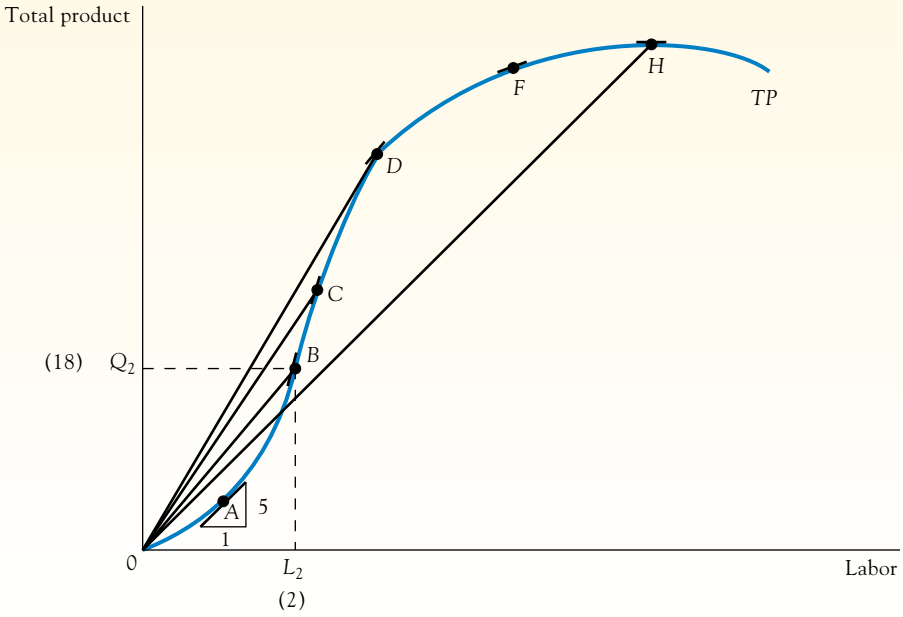
Now consider points *B*, C, and *D* on the total product curve. As output expands from *B* to C to *D*, the straight-line segments 0*B*, 0C, and 0*D* become successively steeper, so that the slopes of these segments become successively greater. This shows that the average product of labor rises over this region. At point *D*, in fact, average product reaches a maximum, because the straight-line segment 0*D* is the steepest such segment from the origin that still touches the total product curve (0*D* is tangent to the total product curve at point *D*). Beyond point *D*, the slope of the straight-line segment connecting the origin to the total product curve begins to decline as the employment of labor is increased. For example, at point *H*, the slope of straight-line segment 0*H* is less than the slope of segment 0*D*. This indicates that the average product of labor is smaller at point *H* than at point *D*.

Marginal product measures how much total output changes with a small change in the use of an input, holding the use of other inputs constant. Figure 7.2 also shows how marginal

FIGURE 7.2

Deriving Average and Marginal Product

The average product of labor equals the slope of a straight-line segment from the origin to any point on the total product curve. Thus, at point *B*, the average product is shown by the slope of straight-line segment 0*B*, or 9 units of output per unit of labor. The marginal product of labor is equal to the slope of the total product curve at each point. Thus, at point *A*, the marginal product is equal to 5 units of output per unit of labor.



product is derived geometrically from the total product curve. *The marginal product of labor at any point on the total product curve is shown by the slope of the total product curve at that point.* The slope of the total product curve is, in turn, equal to the slope of a line tangent to the curve. At point A, for example, we have drawn a line tangent to the total product curve, with a slope of 5/1. Thus, the marginal product of labor at point A is 5 units of output.

The steeper the total product curve, the faster output rises when more input is used, which implies a larger marginal product. In the diagram, marginal product rises as we move up the curve from the origin to point *B*, but it declines (the slope becomes smaller) as we go beyond point *B*. At point *B* the total product curve is steepest, and marginal product is at a maximum. Beyond point *B* output rises less and less when more and more input is used. Note that at point *D* marginal and average product are equal since the slope of the total product curve (marginal product) equals the slope of a straight-line segment from the origin (average product). This is the graphic representation of the proposition, noted earlier, that when average product is at a maximum, marginal and average product are equal. When marginal product falls to zero at point *H*, as implied by the zero slope of the total product curve there, total output is at a maximum.

LAW OF DIMINISHING MARGINAL RETURNS a relationship between output and input that holds that as the amount of some input is increased in equal increments, while technology and other inputs are held constant, the resulting increments in output will decrease in magnitude

The Law of Diminishing Marginal Returns

The shapes of the product curves in Figures 7.1 and 7.2 reflect the **law of diminishing marginal returns**, an empirical generalization about the way output responds to increases in the employment of a variable input. The law of diminishing marginal returns holds that as the

amount of some input is increased in equal increments, while technology and other inputs are held constant, the resulting increments in output will eventually begin to decrease. Put more briefly, the law holds that beyond some point the marginal product of the variable input will decline.

The law of diminishing marginal returns makes intuitive sense. If we begin with 1 worker and 3 units of capital, that worker must be responsible for everything. A second worker may increase total product more than the first worker does if there are advantages to teamwork and the division of labor in producing output. For example, take the case of a firm that delivers pianos to various buyers in a city using the two inputs of trucks (assumed to be fixed at one unit) and labor. A piano is hard for just one worker to move. Two workers working as a team, however, are likely to be more productive than one trying to do the work alone. More teamwork and division of labor are possible as additional workers are employed, but, eventually, the marginal product of additional units of labor falls, because the workers' tasks become redundant and they get in each other's way. Imagine 20 workers crowded around and trying to move a single piano! Ultimately, the marginal product of an extra unit of labor becomes negative when there are so many workers relative to the other, fixed inputs that their efforts actually lower total output. If 20 workers had to squeeze into the firm's one moving truck (the fixed input) there would be little room left for pianos.

In Figure 7.1b, diminishing marginal returns set in when the amount of labor increases beyond two workers. Each additional worker beyond the second adds less to total product than the previous one; the marginal product curve slopes downward. Note that the law of diminishing marginal returns does not depend on workers being different in their productive abilities. We are assuming that all workers are alike.

It is entirely possible that diminishing marginal returns will occur from the very beginning, with the second unit of labor adding less to total output than the first. More commonly, marginal returns increase at very low levels of output and then decline, as in Figure 7.1b. Note also that the law of diminishing returns applies to labor so long as the marginal product of labor curve is downward sloping (as it is beyond two workers). The height of the curve does not have to be negative (as it is beyond eight workers) for the law of diminishing returns to hold.

In applying the law of diminishing marginal returns, two conditions must be kept in mind. First, some other input (or inputs) must stay fixed as the amount of the input in question is varied. The law does not apply, for example, to a situation where labor and capital are the only inputs and the usage of both is increased. It does apply if the amount of capital is held constant while workers and raw materials, for example, are varied. The key point is that some important input is not varied. Second, technology must remain unchanged. A change in technical know-how would cause the entire total product curve to shift.

APPLICATION 7.2

THE LAW OF DIMINISHING MARGINAL RETURNS, CAFFEINE INTAKE, AND EXAM PERFORMANCE

One of the world's most commonly used drugs, caffeine, is a bitter, naturally occurring substance found in coffee and cocoa beans, tea leaves, kola nuts, and other plants.[3] Caffeine is ingested when consuming coffee, tea, soft drinks, or chocolate. Through its ability to stimulate the central nervous system, caffeine can heighten physical performance, mental alertness, and wakefulness—a phenomenon well known to college students at exam time. Research indicates, however, that the law of diminishing returns applies to caffeine

[3]"UCLA Dining Services: The SNAC Guide to Nutrition," *www.dining.ucla.edu*.

consumption. For example, whereas the first cup of coffee may improve the typical student's alertness and thereby her score on an upcoming test, excessive caffeine use results in anxiety, irritability, and trembling. For most students, that is, the 10th cup of coffee is likely to contribute less to their performance on a test than does the ninth cup. And drinking a 10th cup may actually lead to a lower test score on account of the jitters produced by the additional caffeine.

7.3 PRODUCTION WHEN ALL INPUTS ARE VARIABLE[4]

SHORT RUN
a period of time in which changing the employment levels of some inputs is impractical

LONG RUN
a period of time in which the firm can vary all its inputs

VARIABLE INPUTS
all inputs in the long run

By investigating the case where one input (capital) is fixed, the previous section was in fact focusing on the short-run output response by a firm. The **short run** is defined as a period of time in which *changing the employment levels of some inputs is impractical.* By contrast, the **long run** is a period of time in which *the firm can vary all its inputs*. A commercial real estate developer in New York City can acquire the additional land and building permits necessary to supply more office space. DaimlerChrysler has sufficient time to expand its capacity to produce PT Cruisers. There are no fixed inputs in the long run; all inputs are **variable inputs**.

Of course, the distinction between the short run and the long run is necessarily somewhat arbitrary. Six months may be ample time for the clothing industry to make a long-run adjustment to a change in prevailing fashions but insufficient time for the automobile industry to switch from production of large to small cars. Even for a given industry no specific time period can be identified as *the* short run since some inputs may be variable in three months, others in six months, and still others only after a year. Despite this unavoidable imprecision, the concepts of short run and long run do emphasize that quick output changes are likely to be accomplished differently from output changes that can take place over time.

Production Isoquants

ISOQUANT
a curve that shows all the combinations of inputs that, when used in a technologically efficient way, will produce a certain level of output

In the long run all inputs may be varied, so it is necessary to consider all the possibilities identified by the firm's production function. When we consider a product produced by using two inputs, the production options when both inputs can be varied may be shown with isoquants. An **isoquant** is a curve that shows all the combinations of inputs that, when used in a technologically efficient way, will produce a certain level of output. Figure 7.3 shows several isoquants for a firm interested in maximizing output by using the two inputs of capital and labor. Isoquant IQ_{18}, for example, shows the combinations of inputs that will produce 18 units of output. (Note that the axes measure the quantities of the two inputs used.) Combining 5 units of capital with 1 unit of labor will result in 18 units of output (point *B*); so will 2 units of capital and 3 units of labor (point C) or, indeed, any other combination on the IQ_{18} isoquant. Isoquants farther from the origin indicate higher output levels.

The Figure 7.3 isoquants portray an important economic assumption: a firm can produce a particular level of output in various different ways—that is, by using different input combinations, as indicated by points *A*, *B*, *C*, and *D* on IQ_{18}. The firm can produce 18 units of output with a small amount of capital combined with a relatively large amount of labor (point *A*) or with more capital coupled with less labor (point *B*). For example, a car can be custom-built in a local garage with very little equipment and a great deal of labor, or it can be produced in a factory with a large quantity of specialized equipment and far less labor.

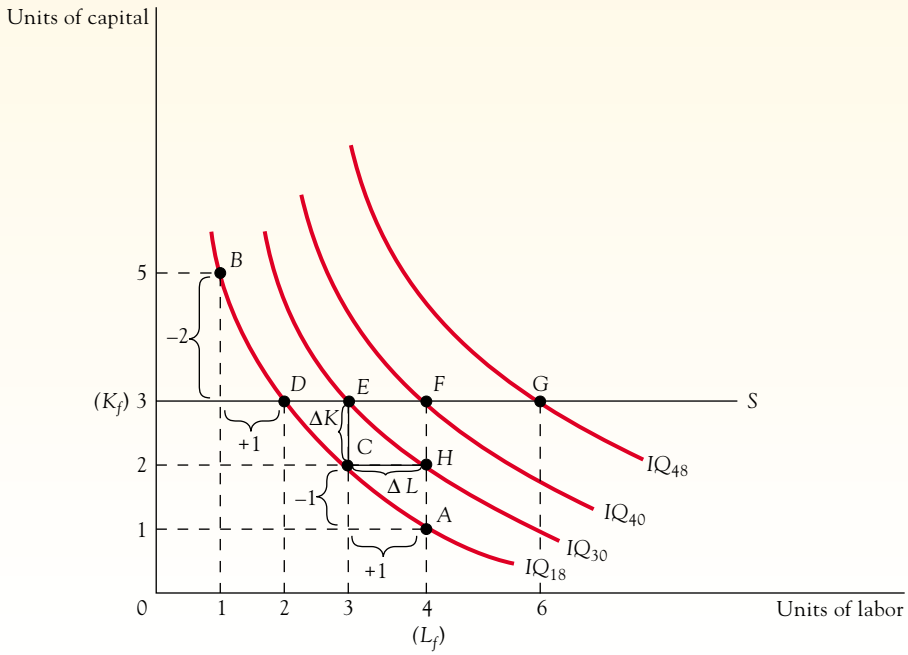
It should be emphasized that *every* combination of inputs shown on the isoquants in Figure 7.3 is technologically efficient: each combination shows the maximum output possible

[4]A mathematical treatment of some of the material in this section is given in the appendix at the back of the book (page xxx).

FIGURE 7.3

Production Isoquants

Production isoquants show how much output a firm can produce with various combinations of inputs. A set of isoquants graphs the production function of the firm. Isoquants have geometric properties that are similar to those of indifference curves: they are downward-sloping, nonintersecting, and convex. The slope of an isoquant measures the marginal rate of technical substitution between the inputs. Between points *B* and *D* the $MRTS_{LK}$ equals $2K/1L$, implying that 1 unit of labor can replace 2 units of capital without reducing the firm's output.



from given inputs. Since a given product can be produced in many different technologically efficient ways, knowing the technological input–output relationships does not by itself allow us to identify the best, or least costly, input combination to use. To determine the lowest-cost way to produce a given level of output, we also need to know input costs, as we will see in the next chapter.

Isoquants are very similar to indifference curves in their characteristics. While indifference curves order levels of a consumer's satisfaction from low to high, isoquants order levels of a producer's output. In contrast to indifference curves, however, each isoquant reflects a measurable output level. As we discussed in Chapter 3, there is no meaningful way to measure the level of satisfaction associated with each indifference curve. The numerical labels associated with each indifference curve are useful only to the extent that they show that higher indifference curves reflect higher levels of consumer satisfaction.

Four characteristics of the isoquants depicted in Figure 7.3 are worth noting. First, the isoquants must slope downward as long as both inputs are productive—that is, they both have positive marginal products. If we increase the amount of labor employed (which, presumably, would by itself raise output) and wish to keep output unchanged, we need to reduce the amount of capital. This relationship implies a negative slope.

Second, isoquants lying farther to the northeast identify greater levels of output. Assuming, again, that inputs are productive, using more of both inputs means a higher output.

Third, two isoquants can never intersect. Intersecting isoquants would imply, at the point of intersection, that the same combination of inputs is capable of producing two different *maximum* levels of output—a logical impossibility.

Fourth, isoquants will generally be convex to the origin. In other words, the slope of an isoquant (in absolute value) becomes smaller as we move down the curve from left to right. To see why this is likely to be true, note that the slope of an isoquant measures the ability of one input to replace another in production. At point B in Figure 7.3, for example, 5 units of capital and 1 unit of labor result in 18 units of output. The input combination at point D, though, can also produce the same output. The slope of the isoquant between B and D is (−2 units of capital)/(+1 unit of labor), meaning that at point B, 1 unit of labor can replace, or substitute for, 2 units of capital without affecting output.

Without the minus sign, the isoquant's slope measures the marginal rate of technical substitution between inputs. The **marginal rate of technical substitution** of labor for capital ($MRTS_{LK}$) is defined as *the amount by which capital can be reduced without changing output when there is a small (unit) increase in the amount of labor*. Between points B and D the $MRTS_{LK}$ is 2 units of capital per 1 unit of labor, which equals the slope when we drop the minus sign.

MARGINAL RATE OF TECHNICAL SUBSTITUTION
the amount by which one input can be reduced without changing output when there is a small (unit) increase in the amount of another input

Convexity of isoquants means that the marginal rate of technical substitution diminishes as we move down each isoquant. Between points C and A on IQ_{18} in Figure 7.3, for example, the $MRTS_{LK}$ is only 1 unit of capital per unit of labor, less than it is between points B and D. The assumption of convexity of isoquants, just as with convexity of indifference curves, is an empirical generalization that cannot be proven correct or incorrect on logical grounds. It does, however, agree with intuition. At point B, capital is relatively abundant, and labor is relatively scarce, compared with point C. Between points B and D, 1 unit of the scarce input (labor) can replace 2 units of the abundant input (capital). Moving down the isoquant, labor becomes more abundant and capital more scarce. It makes sense that it becomes increasingly difficult for labor to replace capital in these circumstances, and this is what is implied by the convexity of the curve.

We have been focusing on the long-run scenario in which a firm can vary the use of all of its inputs, but the isoquants depicted in Figure 7.3 also show that there are diminishing returns to both labor and capital. For example, if we hold capital constant at 3 units (as we did in Section 7.2), and increase the use of labor along line K_fS (K_f signifies that capital is fixed), each additional unit of labor beyond 2 units can be seen to add less and less to output. Increasing labor from 2 to 3 units results in output increasing from 18 to 30 units (from point D on IQ_{18} to point E on IQ_{30} along line K_fS), thus implying that the third worker's marginal product is 12 units of output. Adding a fourth worker raises output from 30 to 40 units (from point E on IQ_{30} to point F on IQ_{40} along line K_fS), indicating that the fourth worker's marginal product is 10 units of output. Since the fourth worker contributes less to output than does the third worker, there are diminishing returns to labor over this range of employment.

The firm also faces diminishing returns to capital. For example, holding labor employment constant at 4 units (L_f where the subscript "f" indicates that labor is being held fixed) and increasing the use of capital from 1 to 2 units raises output from 18 to 30 units (from point A on IQ_{18} to point H on IQ_{30} along segment L_fF). This indicates that the marginal product of the second unit of capital is 12 units of output. Further raising capital from 2 to 3 units, assuming that labor employment is still held constant at L_f, increases output from 30 to 40 units (from point H on IQ_{30} to point F on IQ_{40} along segment L_fF). Since the third unit of capital has a lower marginal product (10 units of output) than does the second unit of capital (12 units of output), the law of diminishing marginal returns applies to capital over this range of capital use.

MRTS and the Marginal Products of Inputs

The degree to which inputs can be substituted for one another, as measured by the marginal rate of technical substitution, is directly linked to the marginal productivities of the inputs. Consider again the *MRTS*, or slope, between points *B* and *D* in Figure 7.3. Between these two points one unit of labor can replace two units of capital, so labor's marginal product must be two times as large as capital's marginal product when the slope of the isoquant ($MRTS_{LK}$) is two units of capital to one unit of labor. To check this reasoning, note that between points C and A in Figure 7.3 the slope of the isoquant is unity. Here the marginal products must be equal because the gain in output from an additional unit of labor (that is, labor's marginal product) must exactly offset the loss in output associated with a 1-unit reduction in capital (that is, capital's marginal product).

Thus, the marginal rate of technical substitution, which is equal to (minus) the slope of an isoquant, is also equal to the relative marginal productivities (*MP*s) of the inputs. Thus:

$$MRTS_{LK} = (-)\,\Delta K/\Delta L = MP_L/MP_K.$$

Note that the isoquant's slope does not tell us the absolute size of either marginal product but only their ratio.

We can also derive this relationship more formally. In Figure 7.3, consider the slope of isoquant IQ_{30} between points *E* and *H*, $\Delta K/\Delta L$. With a move from point *E* to C, the reduction in capital, ΔK, by itself reduces output from 30 to 18 units. This reduction in output must equal ΔK times the marginal product of capital. For example, if $\Delta K = -1$ unit, and the marginal product of the incremental unit of capital is 12 units of output, reducing the amount of capital by 1 unit reduces output by 12 units. Expressing the change in output as ΔQ, we have:

$$\Delta Q = \Delta K \times MP_K.$$

Similarly, when labor increases from point C to *H*, or by ΔL, output increases by ΔL times labor's marginal product:

$$\Delta Q = \Delta L \times MP_L.$$

For a movement along an isoquant, the output decrease from reducing capital must equal the output increase from employing more labor, so the ΔQ terms are equal. The right-hand terms in the two expressions are therefore equal, and, by substitution, we obtain:

$$\Delta K \times MP_K = \Delta L \times MP_L.$$

Then rearranging terms yields the suggested relationship:

$$\Delta K/\Delta L = MP_L/MP_K.$$

Using *MRTS*: Speed Limits and Gasoline Consumption

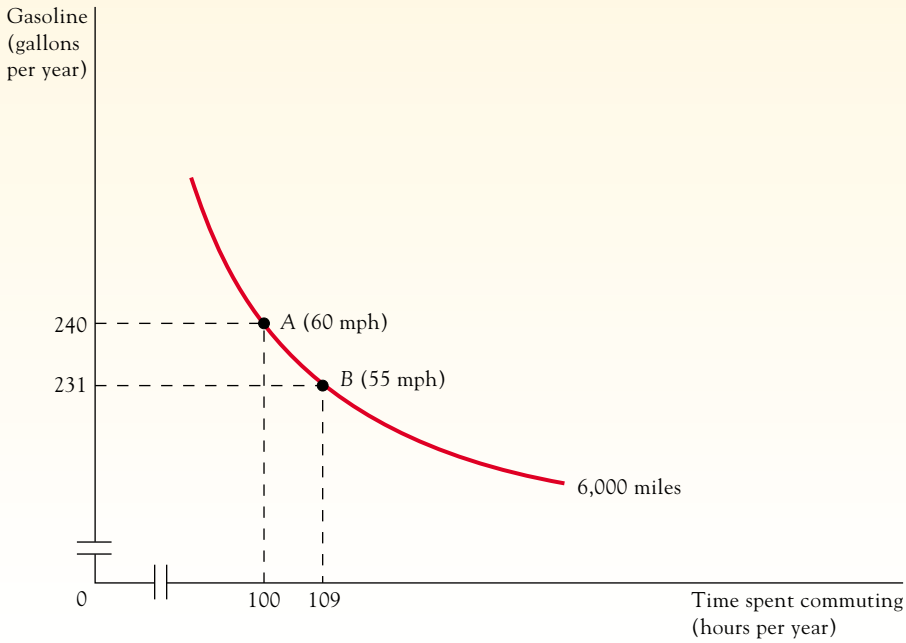
Using isoquants can clarify a wide range of issues. Let's say a person drives 6,000 miles per year to and from work. The speed at which the car is driven affects both the amount of gasoline used (driving faster reduces gas mileage) and the amount of time spent commuting. We can think of gasoline and time as inputs in the production of transportation. Driving slower means using less gasoline but taking more time to get to work. This relationship is shown by the isoquant in Figure 7.4. Suppose that the car gets 25 miles per gallon if driven 60 miles per hour. In that case, commuting at 60 miles per hour uses 240 gallons of gasoline and 100 hours, as shown by point A. If the car gets 26 miles per gallon when driven 55 miles per hour, commuting at 55 miles per hour uses 231 gallons and 109 hours, as shown at point *B*. Driving at the slower speed saves 9 gallons but takes 9 additional hours of commuting time: the *MRTS* is 9 gallons/9 hours, or 1 gallon per hour.

In debates over whether gas savings justify lower speed limits, this isoquant forces us to recognize that there is a tradeoff between gasoline saved by a lower speed limit and addi-

FIGURE 7.4

Isoquant Relating Gasoline and Commuting Time

When driving faster reduces gas mileage, there is a conventionally shaped isoquant relating gas consumption and time. The slope, or *MRTS*, shows the tradeoff between gas and time implied by a change in speed.



tional time spent in transit. The tradeoff is measured by the *MRTS*: here 1 gallon of gasoline per hour spent commuting (between *A* and *B*). Because reducing the speed limit from 60 to 55 miles per hour means using less of one scarce resource (gasoline) but more of another (driver's time), we cannot determine from the *MRTS* alone which speed limit is preferable. Put differently, both *A* and *B* represent technologically efficient points.

Nonetheless, the *MRTS* is one critical piece of information in comparing different speed limits. What else do we need to know? Basically, we need to know the relative importance of the scarce resources, gasoline and time. If gasoline costs $1.00 per gallon, the 55-mile-per-hour speed limit saves our commuter $9.00. But if the commuter values time at anything more than $1.00 an hour (less than one-fifth the minimum wage), the lower speed limit costs the commuter more in lost time than is saved through reduced gasoline use. Another tradeoff is also relevant here: lower speed limits mean greater safety. Once again, the size of the tradeoff between greater safety and time, the *MRTS*, is important. That tradeoff, though, is much harder to measure.

7.4 RETURNS TO SCALE[5]

What relationship exists between output and inputs in the long run? Because all inputs can be varied in the long run, economists approach this problem by focusing on the overall scale of operation. Specifically, we look at how output is affected by a proportionate

[5]A mathematical treatment of some of the material in this section is given in the appendix at the back of the book (pages xxx–xxx).

CONSTANT RETURNS TO SCALE
a situation in which a proportional increase in all inputs increases output in the same proportion

INCREASING RETURNS TO SCALE
a situation in which output increases in greater proportion than input use

DECREASING RETURNS TO SCALE
a situation in which output increases less than proportionally to input use

change in all inputs—for example, when the quantities of both labor and capital are doubled.

In this case three possibilities arise. First, a proportionate increase in all inputs may increase output in the same proportion; for example, doubling all inputs exactly doubles output. Here production is said to be subject to **constant returns to scale**. Second, output may increase in greater proportion than input use: output more than doubles when inputs double. Production is then subject to **increasing returns to scale**. Finally, output may increase less than in proportion to input use. We then have **decreasing returns to scale**.

These are the possibilities, but the actual relationship is not as easy to pin down. Some factors lead to increasing returns and others lead to decreasing returns; which ones predominate in a particular case is an empirical question.

To begin with, what factors may give rise to increasing returns? First, in a large-scale operation workers can specialize in specific tasks and carry them out more proficiently than if they were responsible for a multitude of jobs. This factor, the specialization and division of labor within the firm, was emphasized by the Scottish political economist Adam Smith.

APPLICATION 7.3 ADAM SMITH AND PIN PRODUCTION

In *Wealth of Nations*, Adam Smith noted the increasing returns that division of labor is capable of providing in a business as seemingly simple as the production of pins:[6]

> A workman not educated to this business (which the division of labour has rendered a distinct trade), nor acquainted with the use of the machinery employed in it (to the invention of which the same division of labour has probably given occasion), could scarce, perhaps, with his utmost industry, make one pin in a day, and certainly could not make twenty. But in the way in which the business is now carried on, not only the whole work is a peculiar trade, but it is divided into a number of branches, of which the greater part are likewise peculiar trades. One man draws out the wire, another straights it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make the head requires two or three distinct operations; to put it on, is a peculiar business, to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which, in some manufacturies, are all performed by distinct hands, though in others the same man will sometimes perform two or three of them. I have seen a small manufactory of this kind where ten men only were employed, and where some of them consequently performed two or three distinct operations. But though they were very poor, they could, when they exerted themselves, make among them about twelve pounds of pins in a day. There are in a pound upwards of four thousand pins. . . . Those ten persons, therefore, could make among them upwards of forty-eight thousand pins in a day. Each person, therefore, making a tenth part of the forty-eight thousand pins, might be considered as making four thousand eight hundred pins in a day. But if they had all wrought separately and independently, and without any of them having been educated to this peculiar business, they certainly could not each of them had made twenty, perhaps not one pin in a day; that is, certainly, not the two hundred and

[6]Adam Smith, *The Wealth of Nations* (New York: Modern Library, 1937), pp. 4–5.

fortieth, perhaps not the four thousand eight hundredth part of what they are at present capable of performing in consequence of a proper division and combination of their distinct operations.

As evident in this famous passage, increases in the scale of production in the pin industry allowed firms to realize output increases that were significantly more than in proportion to the increases in input use.

Second, certain arithmetical relationships underlie increasing returns to scale. For example, a 100-foot square building (with 10,000 square feet of floor space) requires 400 feet of walls, but a 100 × 200-foot building, with *twice* the floor space, requires 600 feet of walls, or only 50 percent more material. For another example, a pipeline's circumference (and hence the amount of material that must be employed to create a unit of pipeline) equals the constant "pi" (approximately 3.14) times twice the radius of the pipeline. In contrast, the volume of goods such as crude oil that a pipeline is able to carry depends on the unit area of the pipeline, which equals pi times the pipeline's squared radius. If a pipeline's radius is expanded from 1 to 10 feet, therefore, its circumference (and approximate construction cost) will go up by a factor of 10 while the pipeline's carrying capacity increases by a factor of 100.

Third, the use of some techniques may not be possible in a small-scale operation. Airline hubs, magnetic resonance imaging (MRI) machines, an Internet backbone, assembly lines, direct broadcast satellite television systems, and other similarly complex and expensive techniques or equipment may be feasible only when output is sufficiently high.

The foregoing three factors (division and specialization of labor, arithmetical relationships, and large-scale technologies) are generally what is meant by a phrase such as the "advantages of large-scale or mass production." These factors, however, are inherently limited: after a certain scale of operation is reached, further expansion makes more economies impossible. Even the arithmetical factors may be limited: as a building becomes larger, the ceiling and walls may have to be built with stronger materials; and as a pipeline is enlarged, stronger materials may have to be employed as well as proportionately greater amounts spent on pumping crude oil through the pipeline.

Set against the factors leading to increasing returns to scale is one factor that tends to produce decreasing returns to scale: the inefficiency of managing large operations. With large operations, coordination and control become increasingly difficult. Information may be lost or distorted as it is transmitted from workers to supervisors to middle management and on to senior executives, and the reverse is equally likely. Communication channels become more complex and difficult to monitor. Decisions require more time to make and implement. Problems of this sort occur in all large organizations, and they suggest that the managerial function can be a source of decreasing returns to scale.

APPLICATION 7.4

THE MANAGEMENT FUNCTION AND DECREASING RETURNS TO SCALE: "THE PLAN"

That the managerial function can be a source of decreasing returns to scale is attested to by the following anonymously written and allegorical tale of life within a large corporation:

The Plan

In the beginning was the Plan.
And then came the Assumptions.
And the Assumptions were without form.

And the Plan was without substance.

And darkness was upon the faces of the Workers.

And they spoke among themselves, saying, "It is a crock of s**t, and it stinketh."

And the Workers went unto their Supervisors and said, "It is a pail of dung, and none may abide the odor thereof."

And the Supervisors went unto their Managers, saying, "It is a container of excrement, and it is very strong, such that none may abide by it."

And the Managers went unto their Directors, saying, "It is a vessel of fertilizer, and none may abide its strength."

And the Directors spoke amongst themselves, saying one to another, "It contains that which aids plant growth, and it is very strong."

And the Directors then went unto the Vice Presidents, saying unto them, "It promotes growth, and it is very powerful."

And the Vice Presidents went unto the President, saying unto her, "This new Plan will actively promote the growth and vigor of the company with powerful effects."

And the President looked upon the Plan and saw that it was good.

And the Plan became Policy.

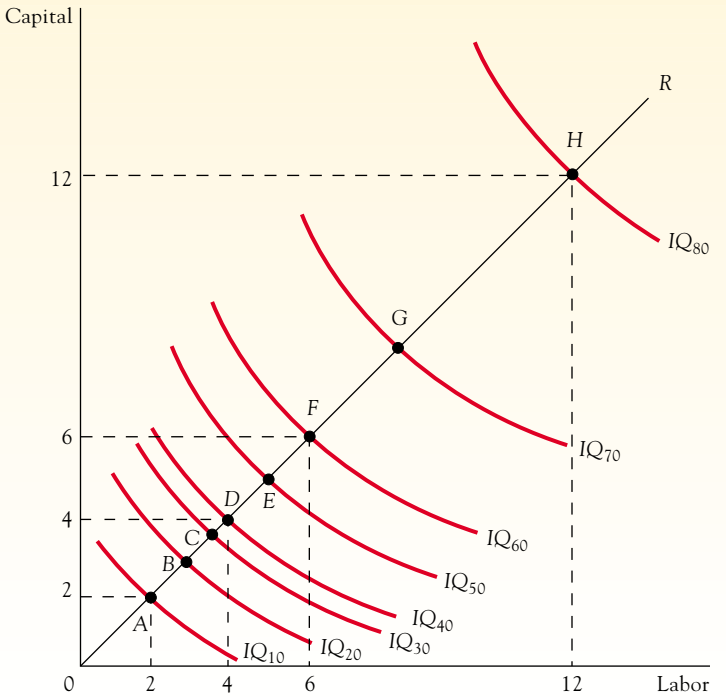
The relative importance of the factors leading to increasing and decreasing returns to scale is likely to vary across industries. As a general rule, increasing returns to scale are likely to apply when the scale of operation is small, perhaps followed by an intermediate range when constant returns prevail, with decreasing returns to scale becoming important for large-scale operations. In other words, a production function can embody increasing, constant, and decreasing returns to scale at different output levels. In fact, this condition is probably the general case.

Figure 7.5 shows isoquants reflecting such a production function. Because we are talking about returns to scale, we are interested in how output varies as we move along a ray from

FIGURE 7.5

Returns to Scale

The spacing of isoquants indicates whether returns to scale are increasing, constant, or decreasing. From *A* to *D*, there are increasing returns to scale; from *D* to *F*, constant returns to scale; and beyond *F*, decreasing returns to scale.



the origin, like 0*R* in the diagram. Along 0*R* the proportion of capital to labor is constant: the ratio of capital to labor is one-to-one at all points. At low output rates, increasing returns to scale prevail: when capital and labor are both doubled, in a move between points *A* and *D*, output more than doubles (that is, it quadruples since we move from IQ_{10} to IQ_{40}). Between points *D* and *F*, a range of constant returns to scale occurs: increasing both capital and labor by a half, in a move from *D* to *F*, increases output by exactly a half (that is, from IQ_{40} to IQ_{60}). Finally, beyond point *F*, decreasing returns to scale result: a doubling of inputs increases output by only one-third (that is, from IQ_{60} to IQ_{80}) in a move from *F* to *H*.

Figure 7.5 shows that as inputs are increased proportionately, the spacing of isoquants provides a graphical method of ascertaining returns to scale. With increasing returns to scale, that is, isoquants become closer and closer to one another as inputs are scaled up proportionately (that is, moving from *A* to *D*). The spacing between isoquants is equidistant with constant returns to scale (moving from *D* to *F*). And the spacing between isoquants grows farther apart with decreasing returns to scale (moving from *F* to *H*).

Of course, saying that returns to scale generally will be increasing at first, then constant, and then decreasing is not saying a great deal. The exact output range over which these relations hold is very important; as we will see in the next chapter, it helps determine the number of firms that can survive in an industry.

APPLICATION 7.5 WHY OIL SHIPPERS ARE COMPARTMENTALIZING THEIR FIRMS AND FLIERS ARE BUILDING THEIR OWN PLANES

Recent oil-spill legislation allows unlimited liability in the case of an accident.[7] Namely, it allows governments to take companies to court for the entire worth of their assets if a spill occurs. Due to the legislation it does not pay to be big if one is in the oil shipping business; the cost of insurance (an important input in the business) increases dramatically with the size of the oil shipping firm. Prior to the imposition of unlimited liability rules, shippers could petition courts to limit liability to the value of the cargo and vessel only. The request was typically approved provided there was not evidence of gross negligence or willful misconduct.

Confronted by legally imposed decreasing returns to scale, oil shippers have been seeking to minimize their insurance costs. Many of the world's largest tanker fleets, often owned by families whose entire fortunes are invested in them, have restructured into a group of smaller firms, each with a tanker as its sole asset. The liability risk (and associated insurance cost) is thereby localized should a spill occur. The strategy is much like that of the builders of the *Titanic*, who divided their ship into many compartments in an attempt to ensure that if one compartment was ever flooded, the other compartments would remain watertight and the ship would stay afloat. Undoubtedly oil shippers are banking on a better end result.

Researchers document a similar phenomenon in manufacturing industries in which there are real or suspected cancer risks.[8] Employing statistical analysis, the researchers found that during the period when liability laws were being rewritten (1967 to 1980), a large increase in the number of small corporations in hazardous sectors occurred. The evidence suggests efforts on the part of corporations to lower legal exposure and the associated insurance costs through divestiture.

The production of small airplanes provides a final example of decreasing returns to scale due to liability issues. The once dominant firm in the industry, Cessna, ceased production of single-engine craft between 1986 and 1995. Another significant supplier, Piper Aircraft, entered bankruptcy in 1991. Both firms cited liability

[7]"Oil Firms, Shippers Seek to Circumvent Laws Setting No Liability Limit for Spills," *Wall Street Journal*, July 26, 1990, pp. B1 and B4.

[8]Al H. Ringleb and Steven N. Wiggins, "Liability and Large-Scale, Long-Term Hazards," *Journal of Political Economy*, 98 No. 3 (June 1990), pp. 574–595.

issues and high insurance costs as the reasons for their actions.[9]

As large producers exited the small plane business, entry took place in the form of individuals building their own planes at home from kits. In recent years, sales of such kits have been roughly twice as large as sales of already-built small planes. According to a 1949 rule, it is legal to fly a plane that has not been certified as airworthy by the Federal Aviation Administration provided that you have built at least half the plane.

We know of at least one former student who has constructed a small plane at home from a kit. He keeps inviting us to drop by for a ride in his self-made craft. So far we have managed to come up with some other plans on the days that he has asked us to visit.

[9]"Liability Costs Drive Small-Plane Business Back Into Pilots' Barns," *Wall Street Journal*, December 11, 1991, pp. A1 and A8. Cessna reentered the market in 1995 when Congress passed a law restricting liability for manufacturers of small aircraft.

7.5 Empirical Estimation of Production Functions[10]

As with demand, production relationships can be estimated through surveys, experimentation, or regression analysis. For example, a former student of ours was employed as a consultant by McKinsey & Company a few summers ago and assigned to a client interested in entering the pig chow (food for pigs) business in certain Midwestern states. One of his tasks was to determine the extent of increasing returns to scale in the production of pig chow. That is, over what range would output continue to go up more than in proportion to the increase in overall input use? The easiest way for the student to determine this range was through telephone surveys of existing producers. Even though most existing producers were reluctant to talk to him regarding the size of their operations for fear of releasing trade secrets, a surprising number provided the relevant data.

Regression analysis offers another method for estimating production functions. Of course, as noted in Chapter 4, such a method is not without its difficulties. Differences in technology across firms must be taken into account. Measures of the amount of each input employed by a firm may not be easy to calculate. In the case of "labor," for instance, most firms employ a wide variety of different types of labor (engineers, clerical assistants, accountants, and so on) at different wage rates. The measurement of output may also involve some difficulties. For example, organizations may not produce a single output, as in the case of universities, which supply both research and teaching services (we discuss multiproduct firms in Chapter 8). Moreover, firms may have access to the same technology but face different regulatory environments. The ability to transform a given amount of inputs into output may be more limited in a restrictive regulatory environment.

In employing regression analysis, care also must be exercised in selecting a functional form for the relationship between inputs and output. Take the case of the following linear relationship between output (Q) and the two inputs of labor (L) and capital (K):

$$Q = a + bL + cK.$$

Such a linear production function is straightforward to interpret and easy to estimate, but presumes that the law of diminishing returns does not apply to either input. To see why, suppose that the estimated intercept and coefficients are $\hat{a} = 0$, $\hat{b} = 4$, and $\hat{c} = 3$ (as noted in Chapter 4, the "$\wedge$" signifies an estimated value). If we start off employing 1 unit of both inputs, the estimated output ($\hat{Q}$) would be 7:

$$\hat{Q} = \hat{a} + \hat{b}(1) + \hat{c}(1) = 0 + 4(1) + 3(1) = 7.$$

[10]A mathematical treatment of some of the material in this section is given in the appendix at the back of the book (pages xxx–xxx).

Fixing the level of capital at 1 unit, and varying the level of labor to 2 units would increase output by 4 units to 11:

$$\hat{Q} = \hat{a} + \hat{b}(2) + \hat{c}(1) = 0 + 4(2) + 3(1) = 11.$$

Upon reaching this output level, varying labor further to 3 would increase output an additional 4 units to 15:

$$\hat{Q} = \hat{a} + \hat{b}(3) + \hat{c}(1) = 0 + 4(3) + 3(1) = 15.$$

And so on. The law of diminishing returns thus can be seen not to apply to labor because each additional unit of labor does not add a diminishing amount but the same amount—4 units—to output. An analogous result applies to capital. Holding fixed the level of labor, say at 1 unit, each additional unit of capital increases output by a constant ($\hat{c} = 3$), rather than a diminishing, amount.

Of course, there are more elaborate mathematical forms of production functions that do not imply constant marginal products for inputs. Among the most common is the **Cobb–Douglas production function**.[11] In the case of our two-input example, the Cobb–Douglas production function takes this form:

COBB–DOUGLAS PRODUCTION FUNCTION
a production function that does not imply constant marginal products for inputs

$$Q = aL^bK^c.$$

Such a multiplicative form allows the law of diminishing returns to either apply or not apply to individual inputs. To see why, suppose that the estimated constant a and powers associated with the inputs labor and capital (b and c, respectively) are $\hat{a} = 2$, $\hat{b} = 0.5$, and $\hat{c} = 1$. If we start off employing 1 unit of both inputs, the estimated output ($\hat{Q}$) would be 2:

$$\hat{Q} = \hat{a}L^{\hat{b}}K^{\hat{c}} = 2(1^{0.5})(1^1) = 2.$$

Fixing the level of capital at 1 unit, and varying the level of labor to 2 units would increase output by 0.83 units to 2.83:

$$\hat{Q} = \hat{a}L^{\hat{b}}K^{\hat{c}} = 2(2^{0.5})(1^1) \approx 2(1.414)(1) \approx 2.83.$$

Upon reaching this output level, varying labor further to 3 units would increase output by 0.63 units from 2.83 to 3.46:

$$\hat{Q} = \hat{a}L^{\hat{b}}K^{\hat{c}} = 2(3^{0.5})(1^1) \approx 2(1.732)(1) \approx 3.46.$$

The law of diminishing returns can thus be seen to apply to labor for the input levels we have considered because, holding constant employment of capital, the third unit of labor adds less to total output (0.63 units) than does the second unit (0.83 units).

In the case of capital, however, the law of diminishing returns does not apply for the assumed Cobb–Douglas production function and estimated constant a and powers b and c. Suppose that we start off once again by employing 1 unit of both inputs. As we have seen before, the estimated output ($\hat{Q}$) is 2:

$$\hat{Q} = \hat{a}L^{\hat{b}}K^{\hat{c}} = 2(1^{0.5})(1^1) = 2.$$

Now instead of holding capital constant, let's fix labor at 1 unit and vary the level of capital to 2 units. Total output would increase by 2 units to 4:

$$\hat{Q} = \hat{a}L^{\hat{b}}K^{\hat{c}} = 2(1^{0.5})(2^1) = 2(1)(2) = 4.$$

[11]This type of production function is named after Charles W. Cobb, a mathematician, and Paul H. Douglas, an economist and U.S. senator. Cobb and Douglas did pioneering work in estimating production functions in the early part of the twentieth century.

Upon reaching this output level, varying capital further to 3 units would increase output by 2 units to 6:

$$\hat{Q} = \hat{a}\hat{L}^{\hat{b}}\hat{K}^{\hat{c}} = 2(1^{0.5})(3^{1}) = 2(1)(3) = 6.$$

The law of diminishing returns thus can be seen not to apply to capital because, holding constant employment of labor, the third unit of capital raises output by the same amount (2 units) as does the second unit of capital.

In general, if the power associated with an input in a Cobb–Douglas production function is less than unity, the law of diminishing returns applies to that input over all possible levels of input usage (do you see why?). If the power associated with an input is equal to or greater than unity, the law of diminishing returns does not apply to that input.

Furthermore, the sum of the powers associated with the inputs in a Cobb–Douglas production function has economic significance. If the sum of the powers exceeds unity (that is, $b + c > 1$), the production function is characterized by increasing returns to scale. If the sum of the powers equals unity ($b + c = 1$), constant returns to scale apply. Decreasing returns to scale apply when the sum of the powers is less than unity ($b + c < 1$).

To see why the sum of the powers associated with the inputs in a Cobb–Douglas production function is related to returns to scale, consider what would happen to output if we scaled up employment of all inputs by some factor, s. The scaling factor s is some number greater than unity because we are contemplating "scaling up" use of all inputs. For example, if we considered doubling all inputs, $s = 2$. To check on returns to scale, we want to compare the output we get when we scale up all inputs by s:

$$a(Ls)^b(Ks)^c,$$

with the original output, Q, scaled up by the same factor s:

$$sQ = saL^bK^c.$$

Written side by side, we are comparing whether the output we get when we scale up all inputs is more than the scaled-up initial output:

$$a(Ls)^b(Ks)^c \text{ versus } saL^bK^c.$$

With some simple rearrangement, the comparison boils down to the following:

$$s^{b+c}aL^bK^c \text{ versus } saL^bK^c.$$

If the sum of the powers associated with the inputs of labor and capital exceeds unity (that is, $b + c > 1$), the above comparison indicates that scaling up all inputs (the left-hand side) will get us more than the scaled-up initial output (the right-hand side). This is the case when increasing returns apply. For example, if the sum of the powers associated with the inputs exceeds unity and $s = 2$, doubling all inputs will get us more than double the initial output.

If the sum of the powers associated with the inputs labor and capital equals unity ($b + c = 1$), scaling up all inputs (the left-hand side of the comparison) will get us the same amount as the scaled-up initial output (the right-hand side). This holds in the case of constant returns to scale. If $s = 2$ in such a case, doubling all inputs will produce an output that is exactly double the initial output.

Finally, if the sum of the powers associated with the inputs labor and capital is less than unity ($b + c < 1$), scaling up all inputs (the left-hand side) will get us less than the scaled-up initial output (the right-hand side) and decreasing returns to scale apply. Were we to double the use of all inputs in such a case (that is, $s = 2$), output would less than double.

SUMMARY

- There are two relationships between the quantities of inputs used and the amount of output produced. In the first, the quantities of some inputs are not changed (fixed inputs), while the quantities of other inputs (variable inputs) are. This is normally a short-run output response, when varying the quantities of some inputs is not practical.
- In the second relationship, the quantities of all inputs can be varied, which is normally the case when long-run output responses are considered.
- With some inputs held fixed, the total product curve shows the relationship between the quantity of the variable input and output.
- The law of diminishing marginal returns holds that beyond some level, the marginal product of the variable input will decline as more of the input is used. This law implies that the total, average, and marginal product curves will have the general shapes shown in Figure 7.1.
- Isoquants depict all input combinations that will produce a given output level. They show the relationship between inputs and output when all inputs can be varied.
- A set of isoquants is effectively a graphical representation of the firm's production function.
- Isoquants and indifferences curves have the same geometric characteristics.
- The marginal rate of technical substitution shows the technological feasibility of trading one input for another and is equal to the slope of an isoquant.
- Returns to scale refer to the relationship between a proportionate change in all inputs and the associated change in output. If output increases in greater proportion than input use, production is said to be subject to increasing returns to scale.
- Constant and decreasing returns to scale are defined analogously. In general, increasing returns to scale are common at low levels of output for a firm, possibly followed by constant returns over a certain range.
- At high levels of output, decreasing returns to scale will exist.
- Although it is not without its difficulties, regression analysis offers one means for estimating the relationship between inputs employed and output.

REVIEW QUESTIONS AND PROBLEMS

Questions and problems marked with an asterisk have solutions given in Answers to Selected Problems at the back of the book (page xxx).

7.1. Fill in the spaces in the accompanying table associated with the firm William Perry, Inc., that delivers refrigerators in the Chicago area, using the two inputs of labor and trucks.

Number of Trucks	Amount of Labor	Total Output	Average Product of Labor	Marginal Product of Labor
2	0	0	—	—
2	1	75		
2	2		100	
2	3			100
2	4	380		
2	5			50
2	6		75	

7.2. State the law of diminishing marginal returns. How is it illustrated by the data in the table of the preceding question? There is a proviso to this law that certain things be held constant: What are these things? Give examples of situations where the law of diminishing marginal returns is not applicable because these "other things" are likely to vary.

***7.3.** If the total product curve is a straight line through the origin, what do the average product and marginal product curves look like? What principle would lead you to expect that the total product curve would never have this shape?

***7.4.** Is it possible that diminishing marginal returns will set in after the very first unit of labor is employed? What do the total, average, and marginal product curves look like in this case?

***7.5.** Deloitte & Touche is thinking of hiring an additional employee. Should the firm be more concerned with the average or the marginal product of the new hire?

7.6. Consider your time spent studying as an input in the production of total points on an economics test. Assume that other inputs (what could they be?) are not varied. Draw the total, average, and marginal product curves. Will they have the general shapes shown in Figure 7.1? Why or why not?

7.7. Define *isoquant*. What is measured on the axes of a diagram with isoquants? What is the relationship between the isoquant map and the production function?

7.8. Assume that the marginal product of each input employed by Microsoft depends only on the quantity of that input employed (and not on the quantities of other inputs), and that

diminishing marginal returns hold for each input. Explain why Microsoft's isoquants must be convex if these assumptions hold.

***7.9.** When United Airlines uses equal amounts of pilots and mechanics, must the isoquant drawn through this point have a slope of -1? Could the isoquant have a slope of -1? If so, what would this characteristic tell us?

7.10. Isoquants are downward-sloping, nonintersecting, convex curves. Explain the basis for each of these characteristics.

***7.11.** For a particular combination of capital and labor we know that the marginal product of capital is 6 units of output and that the marginal rate of technical substitution is 3 units of capital per unit of labor. What is the marginal product of labor?

7.12. Show how a total product curve for an input can be derived from an isoquant map. Why does the question specify "a" total product curve rather than "the" total product curve?

7.13. If the firm's isoquants in Figure 7.3 were straight lines, what would this imply about the two inputs?

7.14. In the commuting example in the text, we assumed that the car in question got 25 miles per gallon if driven at 60 mph and 26 miles per gallon if driven at 55 mph. If the car gets 1 more mile per gallon for each 5-mile-per-hour reduction in speed, will the isoquant be convex? Support your answer by identifying several more points on the isoquant in Figure 7.4.

7.15. Does the concept of technological efficiency permit us to determine at which point on an isoquant a firm should operate?

7.16. Suppose that the number of points on an economics midterm (P) can be characterized by the following production function:

$$P = 15 + 2HB,$$

where H is the number of hours spent studying for the exam and B is the number of beers consumed the week before the exam. Does the law of diminishing returns apply to H? To B? What does the typical isoquant look like for such a production function? Is the production function characterized by increasing, decreasing, or constant returns to scale? Explain your answers.

7.17. Answer all of the questions in the preceding problem if the production function is characterized as follows:

$$P = 5H - 4B.$$

7.18. American Airlines produces round-trip transportation between Dallas and San Jose using three inputs: capital (planes), labor (pilots, flight attendants, and so on), and fuel. Suppose that American's production function has the following Cobb–Douglas form:

$$T = aK^bL^cF^d = 0.02K^{0.25}L^{0.2}F^{0.55},$$

where T is the number of seat-miles produced annually, K is capital, L is labor, and F is fuel.

a. If American currently employs $K = 100$, $L = 500$, and $F = 20{,}000$, calculate the marginal products associated with K, L, and F.

b. What is American's marginal rate of technical substitution ($MRTS$) between K and L? How about the $MRTS$ between K and F? Should American try to ensure that all its $MRTS$s are equal? Explain.

c. Does the law of diminishing returns apply to K in the production of seat-miles between Dallas and San Jose by American? To L or F? Explain. Would the law of diminishing returns apply to L if $c = -0.2$ instead of 0.2? If $c = 1.2$?

d. Given that the exponent associated with F is larger than the exponent associated with L, would it be wise for American to spend all its money on either fuel or capital and none on labor? Explain.

e. Does American's production function exhibit constant, increasing, or decreasing returns to scale? Explain. How would your answer change if $c = -0.2$ instead of 0.2? If $c = 1.2$?

f. Does the law of diminishing returns imply decreasing returns to scale? Explain. Would decreasing returns to scale imply the law of diminishing returns?

g. In the real world, do you think that the production of seat-miles between Dallas and San Jose is characterized by a multiplicative, Cobb–Douglas technology? If not, explain the nature of the production function that might characterize a typical firm producing seat-miles in this city-pair market.

7.19. Economists classify production functions as possessing constant, decreasing, or increasing returns to scale. Yet, from a cause-and-effect point of view, it is not readily apparent why decreasing returns to scale should ever exist. That is, if we duplicate an activity we ought to get duplicate results. Hence, if we truly duplicate all of the inputs, we ought to get double the output. Can you reconcile the apparent contradiction between this logic and the expectation of the economist that beyond certain output ranges firms will confront decreasing returns to scale?

7.20. Suppose that you estimated a production function for various professional tennis players. The measure of output is the percentage of matches played by a player that are won by the player. Inputs include the average number of hours per week spent practicing tennis. Suppose that your results indicate that for the 2000 season, the marginal product associated with practice time is 0.07 for Anna Kournikova, 0.09 for Venus Williams, and 0.16 for Monica Seles. If the law of diminishing marginal returns holds, which of the three players would you say spent the most time practicing during the 2000 season?

7.21. The Los Angeles Lakers were the champions of the National Basketball Association during the 1999–2000 season. Two of the Lakers' leading players, Shaquille O'Neal and Kobe Bryant, made 57 and 44 percent, respectively, of the field goal shots they took on the way to capturing the championship. Given these different marginal products, wouldn't the Lakers have done even better in terms of overall scoring had O'Neal taken more shots and Bryant fewer?

7.22. A fellow student states that it is best to stop studying once you reach the point of diminishing returns with regard to the number of hours spent studying. Assess the validity of her statement.

7.23. Nineteenth-century British economist Thomas Malthus reasoned that because the amount of land is fixed, as population grows and more labor is applied to land, the productivity of labor in food production would decline, leading to widespread famine. This prediction is what led economics to be called the "dismal science." Malthus's prediction failed to come to pass as advances in technology, such as the Green Revolution, greatly increased labor productivity in food production. Do such technological advances contradict the law of diminishing marginal returns?

7.24. In 1965, Gordon Moore, the co-founder of Intel, predicted that the number of transistors per square inch on integrated circuits, and thus the computing speed of a given size microprocessing chip, would continue to double every year for the foreseeable future. In subsequent years the pace has slowed down a bit, but data density has doubled approximately every 18 months. This is the current definition of *Moore's Law*. Does Moore's Law contradict the law of diminishing marginal returns?

7.25. Among the key inputs to a houseplant's success are light, temperature, humidity, soil quality, nutrients, pest control, and water. Explain why increased use of any of the inputs such as water is likely to be subject to the law of diminishing marginal returns.

7.26. In the early days of People Express, the top management team at the airline was personally involved in the training and selection of employees. This participation was key to instilling spirit and dedication among the staff, and the organizational culture that resulted led to the airline's successful initial growth (started in 1981, People Express grew in a few years from 3 to 80 planes, reached both U.S. coasts and Europe, and earned positive profits). Explain why decreasing returns might have set in with continued expansion however, and thus ultimately led to the company's demise in 1986.

7.27. In 1998, Mark McGwire hit 70 home runs while playing for the St. Louis Cardinals. In 1999, McGwire hit 65 home runs. This decrease in marginal (home runs per season) product led to an associated decrease in McGwire's average (home runs per season) product. True, false, or uncertain? Explain.