Managerial Economics

A Problem-Solving Approach

Nick Wilkinson

CAMBRIDGE

PART III
PRODUCTION AND COST ANALYSIS

Part III (Chapters 5–7) is concerned with the analysis of production and cost relationships. A knowledge of these is fundamental to capacity planning in the long run, as well as scheduling and purchasing i n the short run. Managers can then operate the firm efficiently, manipulating the use of inputs appropriately. Chapter 5 is concerned with the nature of production relationships between inputs and outputs, which in turn determine cost relationships with outputs. The latter are examined in detail in Chapter 6, with a particular emphasis on the use of cost–volume–profit analysis. Chapter 7 is concerned with the estimation of cost relationships, and again makes use of the statistical methods explained in Chapter 4, while describing the particular problems associated with measuring cost relationships. The results of empirical studies are also discussed.

5 Production theory

Contents

Objectives

1 To introduce the concept of production and explain its relevance to managerial decision-making.
2 To explain the meaning and significance of different time frames.
3 To describe the different factors of production and explain the concept of the production function.
4 To explain the different concepts of efficiency.
5 To explain the concept of an input-output table and its applications to different time frames and to isoquants.
6 To explain isoquant analysis and its applications in both short-run and long-run situations.
7 To explain how an optimal combination of inputs can be determined in both short-run and long-run situations.
8 To explain the parallels between production theory and consumer theory.
9 To describe different forms of production function and their implications.
10 To explain the concept of returns to scale and its relationship to production functions and empirical studies.
11 To describe and explain relationships between total, average and marginal product, and the different stages of production.
12 To enable students to apply the relevant concepts to solving managerial problems.

5.1 Introduction

In the previous chapters we have seen how firms are usually profit-oriented in terms of their objectives and we have focused on the revenue side of the profit equation by examining demand. We now need to examine the other side of the profit equation by considering costs. However, just as we had to examine consumer theory in order to understand demand, we must now examine production theory before we can understand costs and cost relationships. In doing this we shall see that there are a number of close parallels between consumer theory and production theory; there is a kind of symmetry between them. At the end of the chapter we will consider the importance of production theory for managerial decision-making, the focus of this text.

What is production theory? Essentially it examines the ***physical relationships between inputs and outputs***. By physical relationships we mean relationships in terms of the variables in which inputs and outputs are measured: number of workers, tons of steel, barrels of oil, megawatts of electricity, hectares of land, number of drilling machines, number of automobiles produced and so on. Managers are concerned with these relationships because they want to optimize the production process, in terms of efficiency. Certain important factors are taken as given here: we are not considering what type of product we should be producing, or the determination of how much of it we should be producing. The first question relates to the demand side of the firm's strategy, while the second involves a consideration of both demand and cost, which is examined in pricing. What we are considering is how to produce the product in terms of the implications of using different input combinations at different levels of output. For example, we can produce shoes in factories that make extensive use of automatic machinery and skilled labour in terms of machine operators, or we can produce them in factories employing more labour-intensive methods and unskilled labour. We cannot say that one method or technology is better or more efficient than the other unless we have more information regarding the relationships and costs of the inputs involved.

5.2 Basic terms and definitions

We cannot proceed any further in analysis without defining and explaining some basic terms that will be used extensively throughout the next three chapters. One example has already emerged from the discussion in the previous paragraph, and indeed in previous chapters: the term **efficiency**. As we shall see, there are different ways of defining this concept. However, before moving on to efficiency, there are a number of other terms that need to be considered.

5.2.1 Factors of production

This term refers to ***inputs*** or ***resources***; these terms are used interchangeably in this text. They refer to ***anything used in the production and distribution of goods and services***. When economists use the term **factors of production** they usually classify them into three, or sometimes four, categories: land, labour and capital. Entrepreneurship is sometimes added as a fourth factor. These terms are not self-explanatory so each is now discussed in turn.

a. Land

Land is really a combination of two different factors. First, there is the area of land that is needed to produce the good. This may be agricultural land, factory area, shop space, warehouse space or office space. Second, land relates to all ***natural resources***, that is anything that comes from the surface of the land, underneath it or on top of it. Thus we include minerals, crops, wood, and even water and air, though it may seem strange to refer to these as land.

b. Labour

Labour is the easiest of the factors to understand, the input of labour being measured in number of workers, or more precisely, in number of hours worked. Of course, labour is not homogeneous and manual labour is often divided into unskilled, semi-skilled and skilled categories. Labour also includes administrative and managerial workers, though some empirical studies have omitted this important input.[1] In practice we may wish to distinguish between these different categories of labour, especially if we want to evaluate their different contributions to output, as will be seen.

c. Capital

This term can again be confusing to students. It does not refer to money, or to capital market instruments; rather it refers to capital goods, that is ***plant and machinery***. Like labour, this is a highly heterogeneous category, and in practice we might want to distinguish between different types of capital, again especially if we want to evaluate their different contributions to output. For example, we may want to classify personal computers, photocopying machines, printers, fax machines and coffee machines separately.

d. Entrepreneurship

Entrepreneurship refers to the ***ability to identify and exploit market opportunities***. It therefore includes two separate functions. This input is often not considered in economic analysis; it is really more relevant in long-run situations, and it is notoriously difficult to measure. For one thing it is difficult to separate entrepreneurship from management; top management should be concerned with both the functions of entrepreneurship, if they are truly representing the interests of shareholders.

5.2.2 Production functions

These represent the relationships between inputs and outputs in symbolic or mathematical form. In general terms we can say that any production function can be expressed as:

$$Q = f(X_1, X_2, X_3, \ldots)$$

where Q represents output of a product and $X_1, X_2, X_3, \ldots$ represent the various inputs. This function is often expressed as:

$$Q = f(L, K) \tag{5.1}$$

where L represents the labour input and K represents the capital input. This is obviously a considerable oversimplification since not only can there be more inputs but there can also be more outputs, with a complex relationship between them. A company like Daimler-Chrysler, for example, uses a huge variety of different inputs and produces many outputs, some of which are also inputs. Components like fuel injection units, headlight units, brake discs and so on are both inputs into the final product of automobiles and, at the same time, outputs that are sold separately.

The production function in (5.1) does not imply any particular mathematical form. The significance and implications of mathematical form will be considered in more detail later, but at this stage we can consider some of the basic variations, assuming the general form in (5.1) with two inputs:

1	$Q = aL + bK$	Linear	(5.2)
2	$Q = aL + bK + c$	Linear plus constant	(5.3)
3	$Q = aL + bK + cLK$	Linear plus interaction term	(5.4)
4	$Q = aL^2 + bK^2 + cLK$	Quadratic	(5.5)
5	$Q = aLK + bL^2K + cLK^2 + dL^3K + eLK^3$	Cubic	(5.6)
6	$Q = aL^bK^c$	Power	(5.7)

5.2.3 *Fixed factors*

These are the factors of production that ***cannot be changed in the short run***. This does not mean that they cannot be changed at all; they can be changed in the long run. In practice these factors tend to involve that aspect of land that relates to area of land, and capital equipment. The nature of these factors will vary from firm to firm and industry to industry. It also may be physically possible to change these factors in the short run, for example close down a factory, but it is not economically feasible because of the high costs involved (redundancy payments and so on).

5.2.4 *Variable factors*

These are the converse of the fixed factors, meaning that they are ***inputs that can be varied in both short and long run***. In practice this applies mainly to that part of land that relates to raw materials and to labour. Not all labour may be easily varied however, since salaried staff may have long-term contracts, making it difficult to reduce this input. It may be easier to increase it, but even here job searches can take time, especially for top positions.

5.2.5 The short run

This is again a term that has a different interpretation in economics from other aspects of business, including finance. In finance the short run or short term refers to a period of a year or less. In economics this is not such a useful definition because it does not permit so many generalizations, bearing in mind the large differences between firms in terms of their business environments. It is therefore more useful to define the short run as being ***the period during which at least one factor input is fixed while other inputs are variable***. In practice this will vary from firm to firm and industry to industry according to the circumstances. It also means that a firm might have several short-run time frames as more and more factors become variable. This tends to be ignored in analysis since the same general principles apply to any short-run situation, as long as at least one factor is fixed. Sometimes economists refer to the 'very short run', defined as being the period during which all factors are fixed. Although output cannot be varied under such circumstances, different amounts can be supplied onto the market depending on inventory levels.

5.2.6 The long run

This is the converse of the short run, meaning that it is ***the period during which all factors are variable***. One can now see that all the last four definitions are interdependent. It may seem initially that this circularity is a problem and is not getting us anywhere, but we will see that these definitions permit some very useful analysis. Some economists also refer to the 'very long run', which they define as being the period during which technology can also change. However, this is not a frequently used term, perhaps because technology is changing more quickly now; most economists assume that technology is changeable in the long run but not in the short run.

5.2.7 Scale

This term refers to scale of production or organization. It relates to ***the amount of fixed factors that a firm has***. It follows therefore that a firm cannot change its scale in the short run. A firm's scale determines its **capacity**; this can be defined in various ways, but the simplest is that it refers to the ***maximum output that a firm can produce in the short run***. It is also sometimes defined as the output where a firm's level of average cost starts to rise. This may be easier to measure, since in practice it is very rare for a firm to produce at maximum capacity. Producing at maximum capacity is not usually desirable either, although it might initially seem so, because it is not normally efficient. This brings us to the next definition.

5.2.8 Efficiency

It was mentioned at the beginning of this section that efficiency may also be defined in various ways. The two types that concern us here are **technical efficiency** and **economic efficiency**.

a. Technical efficiency

This means that a firm is producing the ***maximum output from given quantities of inputs***. Any production function assumes that a firm is operating at technical efficiency. It follows from this that a given output may be produced in many ways, each one of which may be technically efficient; in other words, that output is the maximum output that can be produced from each different combination of inputs.

b. Economic efficiency

This involves ***producing a given output at least cost***. This usually involves a unique combination of inputs, the levels of these inputs depending on their substitutability and complementarity, and also on their prices. While this aspect is discussed to some extent in this chapter, it is dealt with in more detail in Chapter 6.

5.2.9 Input-output tables

The relationships between inputs and outputs can be represented in an input-output or production table. Table 5.1 shows such a table for a cubic function relating to Viking Shoes, a company that makes trainers. The outputs, measured in pairs of shoes produced per week, are rounded to the nearest unit. The specific form of the function used is as follows:

$$Q = 4LK + 0.1L^2K + 0.2LK^2 - 0.04L^3K - 0.02LK^3 \tag{5.8}$$

The shaded column relates to the short-run situation where capital input is held constant at three machines. This is examined in more detail in the next

Table 5.1. Viking Shoes: input-output table for cubic function

		Capital input (machines), K							
		1	2	3	4	5	6	7	8
Labour input (workers), L	1	4	9	13	18	23	27	31	35
	2	8	17	27	36	45	54	62	70
	3	12	26	39	53	67	80^B	92	102
	4	16	33	50	68	85	102	117	131
	5	18	38	59	80^C	100^A	119	137	153
	6	20	42	64	87	110	131	150	167
	7	20	43	66	90	113	135	155	171
	8	19	41	64	87	110	131	149	164

section. In the long run, any combination of inputs is feasible, and it is again assumed in the production function (and the table derived from it) that technical efficiency is achieved.

5.3 The short run

In the short run we have seen that at least one factor is fixed. The following analysis assumes a two-factor situation, involving labour and capital, where one factor, capital, is fixed. It is relatively easy to generalize the analysis to apply to situations where there is more than one of either type of factor.

5.3.1 *Production functions and marginal product*

In the previous section it was stated that the production function may take various mathematical forms. As we have seen in Chapter 3 the mathematical form of a function is important because it indicates the way in which the explanatory variables affect the dependent variable; these effects can be seen in particular in terms of the marginal effect and the elasticity. We can now examine these effects as they relate to the various forms of production function described in (5.2) to (5.7). First we need to explain more precisely the economic interpretation of marginal effects in the context of production theory.

A marginal effect is given mathematically by a derivative, or, more precisely in the case of the two-input production function, a partial derivative (obtained by differentiating the production function with respect to one variable, while keeping other variables constant). The economic interpretation of this is a **marginal product**. The marginal product of labour is the ***additional output resulting from using one more unit of labour, while holding the capital input constant***. Likewise the marginal product of capital is the ***additional output resulting from using one more unit of capital, while holding the labour input constant***. These marginal products can thus be expressed mathematically in terms of the following partial derivatives:

$$MP_L = \partial Q/\partial L \text{ and } MP_K = \partial Q/\partial K$$

Expressions for marginal product can now be derived for each of the mathematical forms (5.2) to (5.7). These are shown in Table 5.2, in terms of the

Table 5.2. Production functions and marginal product

Production function	Marginal product (of labour)
$Q = aL + bK$	a
$Q = aL + bK + c$	a
$Q = aL + bK + cLK$	$a + cK$
$Q = aL^2 + bK^2 + cLK$	$2aL + cK$
$Q = aLK + bL^2K + cLK^2 + dL^3K + eLK^3$	$aK + 2bLK + cK^2 + 3dL^2K + eK^3$
$Q = aL^bK^c$	$abL^{b\text{-}1}K^c$

marginal product of labour. The marginal product of capital will have the same general form because of the symmetry of the functions.

The linear production function has constant marginal product, meaning that the marginal product is not affected by the level of either the labour or the capital input. This is not normally a realistic situation and such functions, in spite of their simplicity, are not frequently used. The linear form with an interaction term is more realistic, since the marginal product depends on the level of capital input, but the quadratic function is normally preferable to any linear function since it shows marginal product as depending on the level of both labour and capital inputs. The value of *a* would normally be negative (and *c* positive), meaning that marginal product is declining (linearly) as the labour input increases, because of the law of diminishing returns, which is explained later in this section.

However, the last two production functions are the ones most commonly used in practice as being the most realistic. In both cases, marginal product depends on the level of both labour and capital inputs. The cubic function involves marginal product increasing at first and then declining (a quadratic function, with *b* positive and *d* negative). The power function, often referred to as the **Cobb–Douglas function**, has declining marginal product at all levels of input (assuming *b* is less than 1), but the decline is increasing as the input increases. This is generally more realistic than the linear decline associated with the quadratic model.

The main advantage of this last model, the Cobb–Douglas function, is that it involves constant elasticities. The elasticities in this case represent **output elasticities**; the coefficient *b* refers to the elasticity of output with respect to labour. It means that ***every 1 per cent increase in labour input will increase output by b per cent, assuming that the capital input is held constant***. A similar interpretation applies to the coefficient *c*.

Apart from marginal effects and elasticities, one other important economic interpretation can be derived from the mathematical form of the production function. This relates to the concept of returns to scale. The interpretation of this aspect will be considered in the next section, since it relates to the long run.

5.3.2 Derivation of the short-run input-output table

In the long-run situation considered in the previous section the production function that was used to generate the input-output table in Table 5.1 was a cubic. Let us assume that Viking's capital input is fixed at three machines. We can now compute the short-run production function by substituting the value of *K*=3 into the cubic function given by expression (5.8):

$$Q = 4LK + 0.1L^2K + 0.2LK^2 - 0.04L^3K - 0.02LK^3$$

This gives the following cubic form for the short-run production function:

$$Q = 13.26L + 0.3L^2 - 0.12L^3 \tag{5.9}$$

Table 5.3. Viking Shoes: effects on output of adding more variable input

Labour input, *L*	Total output, *Q*	Marginal product, *MP*	Average product, *AP*
0	0		–
		13	
1	13		13
		14	
2	27		13.5
		12	
3	39		13
		11	
4	50		12.5
		9	
5	59		11.8
		5	
6	64		10.7
		2	
7	66		9.4
		−2	
8	64		8

This can then be used to compute the input-output table for the short run, which is given by the shaded column in Table 5.1. As before, these outputs are rounded to the nearest unit. This table can now be augmented with further information; as well as measuring total output or product we can also record marginal product (*MP*) and average product (*AP*). These can either be computed from the output column or by deriving their mathematical functions, as follows:

$$MP = \partial Q/L = 13.26 + 0.6L - 0.36L^2 \tag{5.10}$$

$$AP = Q/L = 13.26 + 0.3L - 0.12L^2 \tag{5.11}$$

In Table 5.3 the values of *MP* and *AP* are computed from the output column for ease of comparison. Two things should be noted regarding this table:

1 The values of *MP* should correspond to mid-values of *L*. This is because the marginal product is associated with the increase in labour input. Thus when labour is increased from two to three workers, for example, the resulting *MP* of 12 units corresponds to $L = 2.5$. This is important for computing the correct values of *MP* from function (5.10). In this case:

$$MP = 13.26 + 0.6(2.5) - 0.36(2.5)^2 = 12.51$$

The same principle also applies to graphing *MP*, as seen in Figure 5.1.

2 The values of *MP* are given to the nearest unit since they are computed from the total output column. They may therefore not correspond exactly to the values obtained by using function (5.10) because of rounding errors. This does not apply to *AP*, where fractions are given.

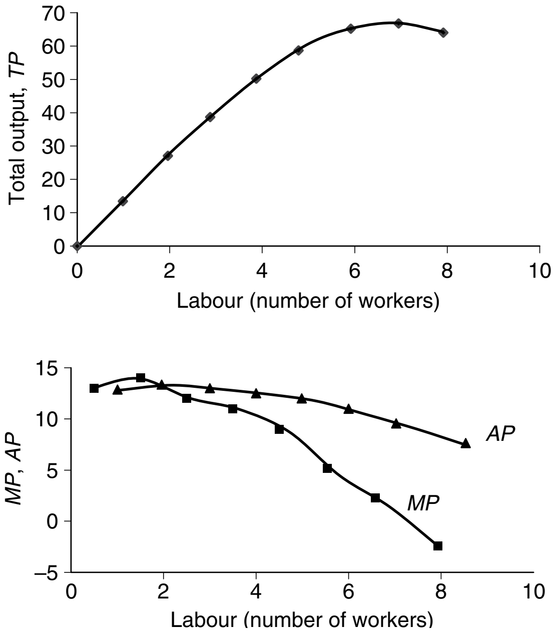

Figure 5.1. Viking Shoes: total, marginal and average product.

5.3.3 *Increasing and diminishing returns*

It can be seen from Table 5.3 that marginal product increases for up to two workers, but when more than two workers are used in combination with the fixed capital input of three machines the marginal product starts to fall. This now needs to be explained.

a. Increasing returns

When the variable input is very low the fixed input, in this case capital, is ***underutilized***; thus one worker cannot use the three machines very efficiently, because he has to do everything himself. When another worker is employed they are able to use the principle of the **division of labour** and specialize in performing different jobs. This increases productivity. As more workers are employed the scope for this increased specialization is reduced, the advantages depending on the amount of capital input and the technology used.

b. Diminishing returns

The **law of diminishing returns** is one of the most important foundations of neoclassical economic theory. It states that ***when additional units of a variable factor are combined with a fixed amount of another factor(s) the additions to total output, in other words the marginal product, will eventually decline***. We have already come across two different applications of this principle: in Chapter 2 the concept of diminishing returns to managerial effort was introduced, and in Chapter 3 the law of diminishing marginal utility was explained. In the first case the fixed factor was the resources of the firm and in the second case the fixed factor was the time frame for consumption. In Table 5.3 the fixed factor is the amount of capital available, three machines. Given that fixed input, and the technology in use, the marginal product of the third worker starts to decline. At this point the fixed factor is becoming ***overutilized***; the workers are beginning to get in each other's way, maybe waiting to use one of the machines. This effect becomes more serious as even more workers are added and marginal output continues to decline, even though total output continues to increase until seven workers are used. When an eighth worker is added the fixed factor becomes so overutilized and workers so crowded together that total output starts to fall; the marginal product now becomes negative.

It must be emphasized that the word *eventually* is important. The law of diminishing returns does not indicate when marginal product will begin to decline. If more of the fixed factor is used, more of the variable factor may also be used in combination with it before the law takes effect, depending on the mathematical form of the production function.

5.3.4 Relationships between total, marginal and average product

Some aspects of these relationships have already been examined in the previous section, but at this point a graphical illustration is helpful. This is given in Figures 5.1 and 5.2. Figure 5.1 is based on the values in Table 5.3 and thus shows the specific pattern for total output, marginal product and average product for Viking Shoes, using the cubic production function in (5.8) and (5.9). Figure 5.2 is more general. It again relates to a cubic form, but not to any specific values; it is shown in order to illustrate various relationships between variable input and total product, marginal product and average product. Figure 5.2 also shows the three stages of the production function, which now need to be explained.

a. Cubic production functions

Three points on the production function or total product (*TP*) curve in Figure 5.2 need to be examined: *A*, *B* and *C*.

Point *A*. It can be seen in Figure 5.2 that the *TP* curve is convex (to the horizontal axis) from the origin to point *A*. As more of the variable input is used the curve now becomes concave, meaning that the slope is decreasing. The slope of the *TP* curve is given by $\partial Q/\partial L$, and this represents the marginal product (*MP*). Thus the *MP* is at a maximum at point *A*, corresponding to the level of input L_1.

Point *B*. The slope of the line, or ray, from the origin to the *TP* curve is given by Q/L and this represents the average product (*AP*). This slope is at a maximum when the line from the origin is tangential to the *TP* curve, at point *B*. This corresponds to the level of input L_2. It can also be seen that this occurs when *AP*

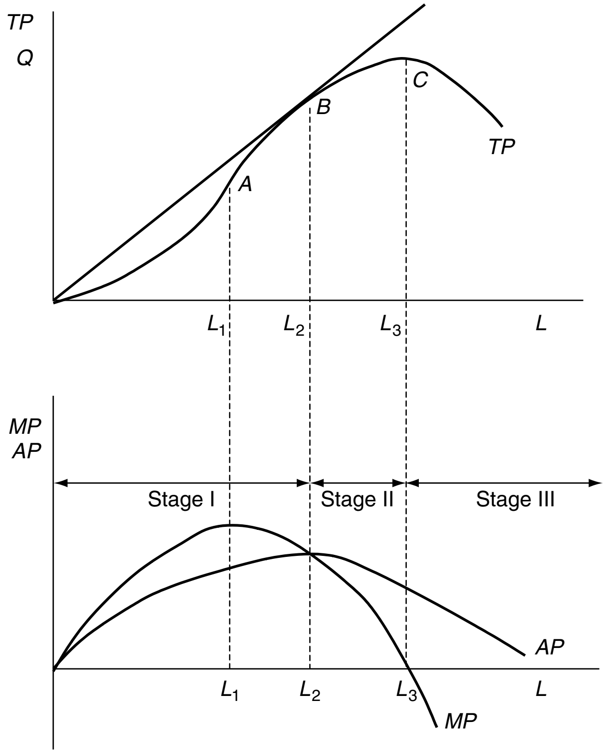

Figure 5.2. Graphical relationships between *TP*, *MP* and *AP*.

and *MP* are equal. This relationship between *AP* and *MP* is important and is paralleled in cost relationships, as we shall see in the next chapter; it therefore needs some explanation. This is most easily done with an analogy to which students can readily relate.

Many students receive a succession of grades or marks on a particular course, assuming continuous assessment is used. It does not matter whether these grades or marks are alphabetical or numerical; they still represent a score. As the course continues, students can compute their average score up to that point. If they wish to improve their average score, their next score, which we can think of as their marginal score, must exceed their existing average. Thus when *MP* is above *AP*, *AP* must be rising (up to input level L_2). If the student's marginal score is less than their average this will pull their average down. Thus when *MP* is below *AP*, *AP* must be falling (above input level L_2). If the marginal score is the same as their average, the average will remain unchanged. Thus when *AP* and *MP* are equal, *AP* remains unchanged and must therefore be a maximum (it is no longer rising and is about to fall).

Point *C*. The *TP* curve reaches a peak or maximum at point *C*; this means that the slope is 0 and therefore *MP* is 0. This corresponds to the level of input L_3. Above this level of input *TP* declines and *MP* becomes negative.

Now that the most important levels of input have been examined we can explain the three different stages of the production function:

- *Stage I*. This corresponds to the input range between zero and L_2, where *AP* reaches a maximum.
- *Stage II*. This corresponds to the input range between L_2 and L_3, where *AP* is falling but *TP* is still rising, meaning that *MP* is still positive.
- *Stage III*. This corresponds to the input range beyond L_3, where *TP* is falling, meaning that *MP* is negative.

The significance of these different stages is that a firm that is operating with economic efficiency will never produce in the stage III region. This is because it is possible to produce the same total output with less of the variable input and therefore less cost. It will be shown in Chapter 8 that, if it is operating in a perfectly competitive environment, the firm should produce in the stage II region in order to maximize profit. Under different demand conditions it is possible for the profit-maximizing output to be in the stage I region.

b. Cobb–Douglas production functions

The shapes of curve shown in Figure 5.2 relate specifically to cubic functions, since these are often found in practice, for reasons explained in the subsection on increasing and diminishing returns. However, the shapes of curve for *TP*, *MP* and *AP* are different for the Cobb–Douglas production function, another function that is found in empirical studies. This function has the form given by (5.7), that is:

$$Q = aL^bK^c$$

Since the output elasticities with respect to labour and capital are both positive the *TP* curve is a continually increasing function.

It was seen in Table 5.2 that the *MP* for this function is given by:

$$MP = abL^{b-1}K^c \tag{5.12}$$

and:

$$AP = Q/L = aL^{b-1}K^c \tag{5.13}$$

If it is assumed that the output elasticity with respect to labour (b) is less than 1, it can be seen from these expressions that both *MP* and *AP* are continually decreasing functions, and that *MP* will be below *AP*. The relevant curves in this case are shown in Figure 5.3.

5.3.5 *Determining the optimal use of the variable input*

Assuming that the capital input, or indeed any input in general terms, is fixed, a firm can determine the optimal amount of the variable input to employ if it uses information relating to product prices and factor costs. This involves an explanation of the concepts of **marginal revenue product** (MRP) and **marginal factor cost** (MFC).

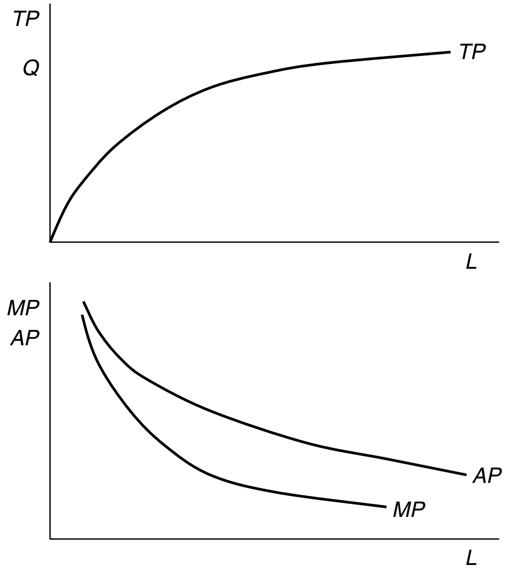

Figure 5.3. Graphical relationships for Cobb–Douglas production functions.

a. Marginal revenue product

This is defined as ***the addition to total revenue from using an additional unit of the variable factor***, or:

$$MRP_L = \Delta R/\Delta L \text{ or } \partial R/\partial L \tag{5.14}$$

assuming that labour is the variable input. This change in total revenue will equal the marginal product of labour times the marginal revenue from selling the additional units of output:

$$MRP_L = (MP_L)(MR_Q) \tag{5.15}$$

The marginal revenue from additional units will be constant if the firm is operating under the conditions of perfect competition, examined in detail in Chapter 8. Let us assume that the selling price for Viking shoes is £75. We can now incorporate this information in Table 5.4 in computing MRP_L.

b. Marginal factor cost

This is defined as ***the addition to total cost from using an additional unit of the variable factor***, or:

$$MFC_L = \Delta C/\Delta L \text{ or } \partial C/\partial L \tag{5.16}$$

Let us assume that the cost of labour is £400 per week. It is also assumed that the firm is operating in a perfectly competitive labour market, meaning that it can employ as many workers as it wants at this going market price. Thus the MFC_L is constant at all input, and therefore output, levels.

Table 5.4. Viking shoes: marginal revenue product and marginal factor cost

Labour input, L (number of workers)	Total output, Q (pairs of shoes)	Marginal product, MP_L(shoes per worker)	Total revenue, $TR = P \times Q$ (£)	Total labour cost, $TLC = L \times MFC_L$ (£)	Total profit, $\Pi = TR - TC$ (£)	MRP_L (£)	MFC_L (£)
0	0		0	0	−1500		
		13				975	400
1	13		975	400	−925		
		14				1,050	400
2	27		2,025	800	−275		
		12				900	400
3	39		2,925	1,200	225		
		11				825	400
4	50		3,750	1,600	650		
		9				675	400
5	**59**		**4,425**	**2,000**	**925**		
		5				375	400
6	64		4,800	2,400	900		
		2				150	400
7	66		4,950	2,800	650		
		−2				−150	400
8	64		4,800	3,200	100		

c. *Profit maximization*

We can now combine the information relating to marginal revenue product and marginal factor cost to determine the profit-maximizing level of use of the variable factor. This will be achieved by expanding the operation as long as the marginal benefits exceed the marginal costs; thus the optimal level of input use is given by the condition:

$$MRP_L = MFC_L \tag{5.17}$$

It can be seen from Table 5.4 that this occurs when the labour input is five workers. Up to this level $MRP_L > MFC_L$, meaning that the additional workers are adding more to revenue than to costs. If more workers are added beyond this level $MRP_L < MFC_L$, the marginal revenue product of the sixth worker is only £375 whereas the marginal factor cost is £400. Thus profit is reduced from £925 to £900. It should be noted that the profit is calculated by subtracting the total cost from total revenue, where the total cost equals the total labour cost plus the cost of using the capital input, assumed in this case to be £1,500 (three machines at £500 each). The profit function can also be expressed mathematically:

$$\Pi = R - C = PQ - C = 75(13.26L + 0.3L^2 - 0.12L^3) - (1500 + 400L)$$
$$\Pi = -1500 + 594.5L + 22.5L^2 - 9L^3 \tag{5.18}$$

The situation is illustrated graphically in Figure 5.4. If the mathematical approach is used, the optimal value of L can be calculated by differentiating the profit function with respect to L and setting the derivative equal to zero.

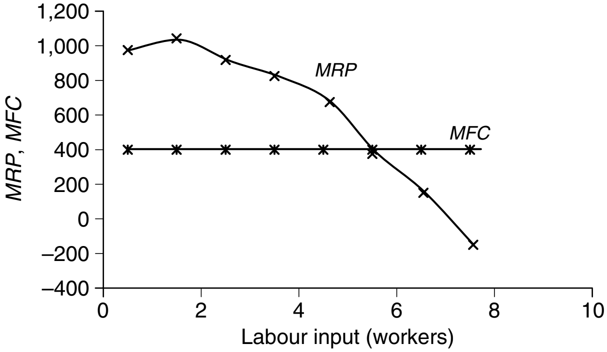

Figure 5.4. Viking Shoes: marginal revenue product and marginal factor cost.

$$\frac{d\Pi}{dL} = 594.5 + 45L - 27L^2 = 0$$

This expression requires solving a quadratic equation. The reader may recall that for the general equation $ax^2 + bx + c = 0$, the solutions for x are given by:

$$x = \frac{-b \pm \sqrt{(b^2 - 4ac)}}{2a}$$

Thus $L = -45 \pm \frac{\sqrt{[45^2 - 4(-27)(594.5)]}}{2(-27)} = 5.60$ (the other solution is negative).

Now that the relevant principles have been discussed as far as the short run is concerned we can turn our attention to a couple of case studies that examine the relevance and application of the law of diminishing returns in real-life situations.

Case study 5.1: Microsoft – increasing or diminishing returns?

In some industries, securing the adoption of an industry standard that is favourable to one's own product is an enormous advantage. It can involve marketing efforts that grow more productive the larger the product's market share. Microsoft's Windows is an excellent example.[2] The more customers adopt Windows, the more applications are introduced by independent software developers, and the more applications that are introduced the greater the chance for further adoptions. With other products the market can quickly exhibit diminishing returns to promotional expenditure, as it becomes saturated. However, with the adoption of new industry standards, or a new technology, increasing returns can persist.[3] Microsoft is therefore willing to spend huge amounts on promotion and marketing to gain this advantage and dominate the industry. Many would claim that this is a restrictive practice, and that this has justified the recent anti-trust suit against the company. The competitive aspects of this situation will be examined in Chapter 12, but at this point there is another side to the situation regarding returns that should be considered.

Microsoft introduced Office 2000, a program that includes Word, Excel, PowerPoint and Access, to general retail customers in December 1999. It represented a considerable advance over the previous package, Office 97, by allowing much more interaction with the Internet. It also allows easier collaborative work for firms using an intranet. Thus many larger firms have been willing to buy upgrades and pay the price of around $230.

However, there is limited scope for users to take advantage of these improvements. Office 97 was already so full of features that most customers could not begin to exhaust its possibilities. It has been estimated that with Word 97 even adventurous users were unlikely to use more than a quarter of all its capabilities. In this respect Microsoft is a victim of the law of diminishing returns.[4] Smaller businesses and home users may not be too impressed with the further capabilities of Office 2000. Given the enormous costs of developing upgrades to the package, the question is where does Microsoft go from here. It is speculated that the next version, Office 2003, may incorporate a speech-recognition program, making keyboard and mouse redundant. At the moment such programs require a considerable investment in time and effort from the user to train the computer to interpret their commands accurately, as well as the considerable investment by the software producer in developing the package.

Questions

1 Is it possible for a firm to experience both increasing and diminishing returns at the same time?
2 What other firms, in other industries, might be in similar situations to Microsoft, and in what respects?
3 What is the nature of the fixed factor that is causing the law of diminishing returns in Microsoft's case?
4 Are there any ways in which Microsoft can reduce the undesirable effects of the law of diminishing returns?

Case study 5.2: State spending

A particularly controversial example of the law of diminishing returns is in the area of state, or public, spending. Some recent studies indicate that diminishing returns have been very much in evidence in developed countries in recent decades, with returns even being negative in some cases. An example is the IMF paper by Tanzi and Schuknecht,[5] which examined the growth in public spending in industrial economies over the past 125 years and assessed its social and economic benefits.

At the beginning of this period, 1870, governments confined themselves to a limited number of activities, such as defence and law and order. Public spending was only an average of 8% of GDP in these countries at this time. The higher taxes that were introduced to pay for the First World War allowed governments to maintain higher spending afterwards. Public spending rose to an average of 15% of GDP by 1920. This spending increased again in the years after 1932 in the surge of welfare spending to combat the Great Depression. By 1937 the average for industrial countries had reached nearly 21% of GDP.

The three decades after the Second World War witnessed the largest increase in public spending, mainly reflecting the expansion of the welfare state. By 1980 the proportion of GDP accounted for by state spending was 43% in industrial countries, and by 1994 this had risen to 47%. By this time there were large variations between countries: the EU average was 52%, in the UK it was 43%, in the USA 33%. In the newly industrializing countries (NICs) the average was only 18%. These variations over time and area allow some interesting comparisons regarding the benefits of additional spending.

Tanzi and Schuknecht found that before 1960 increased public spending was associated with considerable improvements in social welfare, such as in infant-mortality rates, life expectancy, equality and schooling. However, since then, further increases in public spending have delivered much smaller social gains, and those countries where spending has risen most have not performed any better in social or economic terms than those whose spending has increased least. In the higher-spending countries there is much evidence of 'revenue churning': this means that money taken from people in taxes is often returned to the same people in terms of benefits. Thus middle-income families may find their taxes returned to them in child benefits. Furthermore, in many of those countries with the lowest increase in public spending since 1960, efficiency and innovation appear to be greater; they have lower unemployment and a higher level of registered patents.

Another study found a similar pattern in Canada specifically.[6] In the 1960s public spending, at modest levels, helped the development of Atlantic Canada. Most of the money went into genuinely needed roads, education and other infrastructure. Later large increases in spending not only had a smaller effect, but in general had a negative effect. For example, generous unemployment insurance reduced the supply of labour and impeded private investment. Subsidized industries, like coal, steel and fishing, involved using labour that could have been employed in more productive areas, as well as in the last case decimating the cod stocks. Even the roads eventually deteriorated, as local politicians had little incentive to spend public funds wisely, and voters felt unable to discipline them.

Questions

1 In what areas of public spending do there appear to be increasing returns?
2 In what areas of public spending do there appear to be diminishing or negative returns?
3 Explain the difference between diminishing and negative returns in the context of public spending, giving examples.
4 Explain what is meant by 'revenue churning', giving examples.
5 Why do local politicians have little incentive to spend public money wisely?
6 Is it possible to talk about an optimal level of public spending? How might this level be determined?

5.4 The long run

The analysis so far has assumed that at least one factor is fixed, and in the example of Viking Shoes this factor has been capital, being fixed at three machines. We now need to consider the situation where the firm can vary both of its inputs. This means that we need to examine the input-output data in Table 5.1 in more general terms, with both factors being considered as variable. The data in Table 5.1 can also be represented in a three-dimensional graph, but these are generally not very useful for analytical purposes. In order to proceed with any analysis it is necessary to introduce the concept of **isoquants**.

5.4.1 Isoquants

An isoquant is a curve that shows ***various input combinations that yield the same total quantity of output***. It is assumed that the output involved is the maximum that can be produced from those combinations of inputs. Thus the position or equation of an isoquant can be derived from the production function. It corresponds to the concept of an indifference curve in consumer theory, and has analogous properties. For example we can talk of an **isoquant map**, where each curve represents a greater quantity of output as one moves further away from the origin.

The three main properties that isoquants have in common with indifference curves are:

1 *Negative slope.* This is because the inputs are usually assumed to be substitutable for each other; if a firm uses more of one input it needs less of another.
2 *Convexity.* This means that their slope is decreasing from left to right; the reason for this relates to the properties of the **marginal rate of technical substitution**, explained shortly.

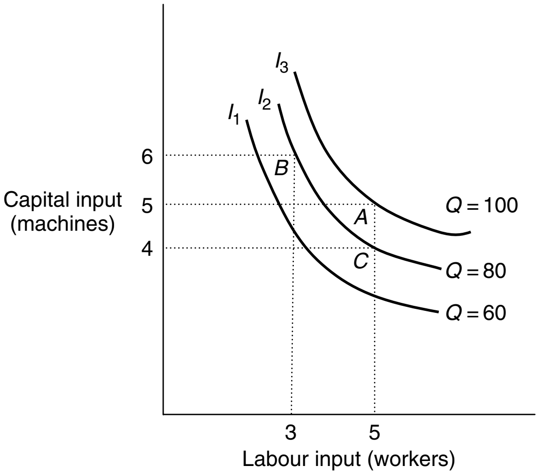

Figure 5.5. Viking Shoes: isoquant map.

3 *Non-intersection.* It is technically possible for isoquants to intersect, as will be seen in the next section, but this will not occur in the economically feasible range of output. If curves intersect it means that a certain output is being produced using more of both inputs, and this is obviously not efficient in economic terms.

Figure 5.5 shows an isoquant map, based on the data in Table 5.1. Points *A, B* and *C* correspond to the values indicated in the table. Thus it can be seen that the output of 80 units can be achieved by using either six machines and three workers (point *B*) or four machines and five workers (point *C*). On the other hand, in order to produce 100 units of output it is necessary to use five machines and five workers (point *A*), though other combinations (involving fractions of inputs) can also produce the same output. It should be noted that the isoquant for the output of 100 units starts to curve upwards as more than seven workers are used; this is because it is not possible to produce 100 units with less than five machines. The maximum output from using only four machines is 90 units, no matter how much labour is used.

5.4.2 The marginal rate of technical substitution

The marginal rate of technical substitution (*MRTS*) is a measure of the degree of substitutability between two inputs. More specifically, ***the MRTS of X for Y corresponds to the rate at which one input (X) can be substituted for another (Y), while maintaining total output constant.*** It is shown by the absolute value of the slope of the isoquant; thus in moving from point *B* to point *C* the *MRTS* is 1, meaning that if two more workers are used we can give up two machines and still produce 80 units of output. The slope of the isoquant is decreasing in absolute magnitude from left to right. This means that as more and more labour is used to produce a given output, the less easily the capital input can be substituted for it. The reason for this is the occurrence of the **law of diminishing returns**, explained in the previous section. Thus as more labour is used and less capital, the marginal product of additional labour falls and the marginal product of the capital lost increases. Relating this to Viking Shoes, it means that as less and less machinery is used it becomes harder to produce a given output with increasing amounts of labour.

At this stage another parallel with consumer theory can be seen: in that case the slope of the indifference curve was shown by the marginal rate of substitution (*MRS*). This was also decreasing in absolute magnitude from left to right, because of the law of diminishing marginal utility.

It was also seen that the *MRS* was given by the ratio of the marginal utilities of the two products. It should not be too difficult for the reader to draw another parallel at this point: the *MRTS* is given by the ratio of the marginal products of the two inputs. The mathematical proof of this is analogous to the one relating to the *MRS*.

When the firm moves from point *B* to point *C* it gains output from using more labour, given by $\Delta L \times MP_L$, and it loses output from using less capital, given by $\Delta K \times MP_K$. Since the points are on the same isoquant and therefore must involve the same total output, the gains must equal the losses, thus:

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

Since the slope of the isoquant is given by $\Delta K/\Delta L$, we can now express the absolute magnitude of the slope as:

$$\Delta K/\Delta L = MP_L/MP_K \tag{5.19}$$

There are two extreme cases of input substitutability. **Zero substitutability** occurs when the inputs are used in fixed proportions, for example when a machine requires two workers to operate it and cannot be operated with more or less than this number of workers. Isoquants in this case are L-shaped, meaning that the *MRTS* is either zero or infinity. **Perfect substitutability** is the opposite extreme, resulting in linear isoquants; this means that the *MRTS* is constant. It also implies that output can be produced using entirely one input or the other. These extremes are shown in Figure 5.6.

5.4.3 Returns to scale

We frequently want to analyse the effects on output of an increase in the scale of production. An increase in scale involves a **proportionate increase in all the inputs** of the firm. The resulting proportionate increase in output determines the **physical returns to scale** for the firm. Two points need to be explained before moving on to the description and measurement of returns to scale:

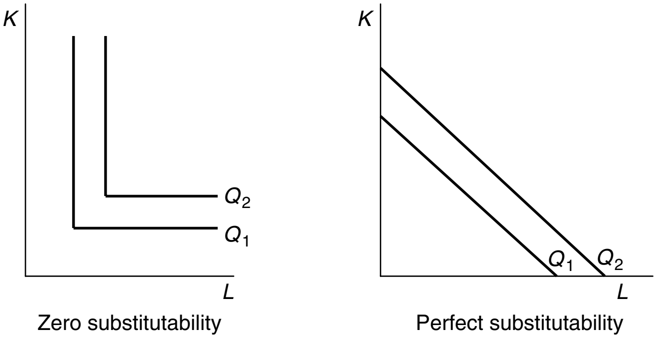

Figure 5.6. Extreme cases of input substitutability.

1 *Proportionate increase in all the inputs*. It is always assumed in referring to returns to scale that all inputs increase by the same proportion. This is not necessarily optimal for the firm in terms of economic efficiency. If inputs increase by different proportions we have to talk about **returns to outlay** (measured in money terms).

2 *Physical returns to scale*. Returns to scale can be described in physical terms or in money terms, as will become clear in the next chapter. The two meanings do not necessarily coincide; for example, it is possible for a firm to experience constant physical returns to scale yet have increasing returns to scale in money terms (better known as economies of scale).

a. Types of returns to scale

Returns to scale, in physical or money terms, can be of three types. The following are the three types of physical return:

1 *Constant returns to scale (CRTS)*. This refers to the situation where ***an increase in inputs results in an exactly proportional increase in output***.

2 *Increasing returns to scale (IRTS)*. This refers to the situation where ***an increase in inputs results in a more-than-proportional increase in output***.

3 *Decreasing returns to scale (DRTS)*. This refers to the situation where ***an increase in inputs results in a less-than-proportional increase in output***.

The reasons for these different returns to scale will be considered in the next chapter, when they are compared with the monetary aspects of returns to scale. We can, however, use Table 5.1 to examine these different possibilities from the standpoint of quantitative measurement. The easiest way to do this is by examining the numbers in the leading diagonal. When inputs are increased from one worker/one machine to two workers/two machines this represents a doubling of inputs; however, output increases from 4 to 17 units, an increase of more than fourfold. Thus this situation involves, IRTS. If inputs increase from two of each factor to three of each factor this is an increase of 50 per cent; output increases from 17 to 39 units, over 100 per cent. Thus there are still

IRTS. This situation continues until seven units of each input are used; when each input is increased to eight units this represents an increase of about 14 per cent, while output increases from 155 to 164 units, an increase of less than 6 per cent. Thus there are now DRTS.

Generalizing from this we can conclude that with a cubic production function the returns to scale are not the same at all levels of scale or output. The type or pattern of returns to scale will obviously depend on the nature of the mathematical form of the production function. In order to understand this more clearly we need to consider the concept of a **homogeneous production function**.

*b. Homogeneous production functions**

These functions are useful for modelling production situations because of their mathematical properties. If the inputs in a function are multiplied by any constant λ and if this constant can then be factored out of the function then the production function is said to be homogeneous. This can be explained more precisely in mathematical terms by stating that a production function is said to be homogeneous of degree n if:

$$f(\lambda L,\ \lambda K) = \lambda^n f(L, K) \tag{5.20}$$

If the degree of homogeneity is equal to 1 then the production function is said to be **linearly homogeneous**. The degree of homogeneity indicates the type of returns to scale:

if $n = 1$ there are CRTS

if $n > 1$ there are IRTS

if $n < 1$ there are DRTS.

These concepts now need to be applied to particular forms of production function. Let us take the simple **linear** form in (5.2) first:

$$Q = aL + bK$$

When each input is multiplied by λ, output is given by:

$$a(\lambda L) + b(\lambda K) = \lambda(aL + bK)$$

Thus λ can be factored out of the function and the function is linearly homogeneous. This means that linear production functions like (5.2) feature constant returns to scale at all levels of output. This is not true for the linear function with a constant term in (5.3); this is not a homogeneous function. Nor is the linear function with an interaction term in (5.4).

Now let us consider the **quadratic** function in (5.5):

$$Q = aL^2 + bK^2 + cLK$$

When inputs are multiplied by λ, output is given by:

$$a(\lambda L)^2 + b(\lambda K)^2 + c(\lambda L)(\lambda K) = \lambda^2(aL^2 + bK^2 + cLK)$$

The quadratic function is also homogeneous, but of the second degree; therefore, there are increasing returns to scale in this case.

Let us now consider the **cubic** function in (5.6):

$$Q = aLK + bL^2K + cLK^2 + dL^3K + eLK^3$$

This function is not homogeneous since the first term will be multiplied by λ^2, the next two terms will be multiplied by λ^3 and the last two terms will be multiplied by λ^4. Since the first three terms are generally positive while the last two are negative we cannot say anything about the type of returns to scale in general. As we have already seen with the cubic function in (5.8), there are increasing returns to scale to begin with and then decreasing returns.

c. *Cobb–Douglas production functions*

Finally let us consider the Cobb–Douglas production function in (5.7):

$$Q = aL^bK^c$$

When inputs are both increased by λ, the resulting output is given by:

$$a(\lambda L)^b(\lambda K)^c = \lambda^{b+c}(aL^bK^c)$$

Thus this type of production function featuring constant output elasticities is homogeneous of order $(b+c)$. This in turn tells us about the type of returns to scale that will occur; ***any increase in inputs of 1 per cent will increase output by (b + c) per cent:***

1 If $b + c = 1$ there are CRTS: a 1 per cent increase in inputs will increase output by 1 per cent.
2 If $b + c > 1$ there are IRTS: a 1 per cent increase in inputs will increase output by $>$ 1 per cent.
3 If $b + c < 1$ there are DRTS: a 1 per cent increase in inputs will increase output by $<$ 1 per cent.

Cobb–Douglas production functions are very useful in practice because of the information they reveal regarding the type of returns to scale in a firm or industry. Empirical findings relating to this aspect will be discussed in Chapter 7, because of their implications regarding costs.

5.4.4 *Determining the optimal combination of inputs*

The isoquants that were considered in the previous analysis all assume that the firm is producing with technical efficiency, which, as we have seen, means that the outputs involved are assumed to be the maximum that could be produced from the combinations of inputs employed. However, for each isoquant there is only one combination of inputs that is economically efficient, meaning minimizing cost, given a set of input prices. The determination of this input combination requires information regarding both the production function, determining the relevant isoquant, and the prices of the inputs employed. This involves moving into aspects of cost analysis, the subject of the next chapter, but there is a difference of perspective. At this point it is assumed that there is a target level of output that is given. The next chapter focuses more on relationships between costs and output where output is treated as a variable.

a. Isocost lines

The prices of the inputs can be used to compute an **isocost** line. This line shows the ***different combinations of inputs that can be employed given a certain level of cost outlay***. We can now see that an isocost line corresponds to the concept of a budget line in consumer theory. Thus the slope of the isocost line is given by the ratio of the input prices, P_L/P_K. Likewise we can derive the firm's optimal position in the same way that we derived the consumer's optimal position.

Let us at this point review the concept of the consumer's optimal position or equilibrium, since it will shed light on the similarities of, and differences between, the optimization procedures involved. In consumer theory the objective was to maximize total utility subject to a budget constraint. The objective that we are now considering in production theory is to minimize cost subject to an output constraint, meaning that we have to produce a certain output. This is called the **dual** of the problem in consumer theory; this corresponds to a kind of mirror image. The differences are as follows:

1 The objective is one of minimization rather than maximization.
2 The isoquants represent maximum outputs that are constraints in production theory, whereas indifference curves represent utilities that are to be maximized in consumer theory.
3 Isocost lines represent costs that are to be minimized in production theory, whereas budget lines represent budgets that are constraints in consumer theory.

In spite of dealing with the 'mirror image' of the problem in consumer theory we can essentially use the same technique of analysis. This can be seen in the graph in Figure 5.7. The optimal point is where the isoquant is tangential to the lowest isocost curve. This is the 'mirror image' of the optimal point in consumer theory, where the budget line is tangential to the highest indifference curve.

It is assumed in the example of Viking Shoes that labour costs £400 per worker per week and capital costs £500 per machine per week. The isocost line C_1 represents a total cost of £3,000 per week, C_2 represents £4,000 per week and C_3 represents £5,000 per week. It can be seen from the graph that the

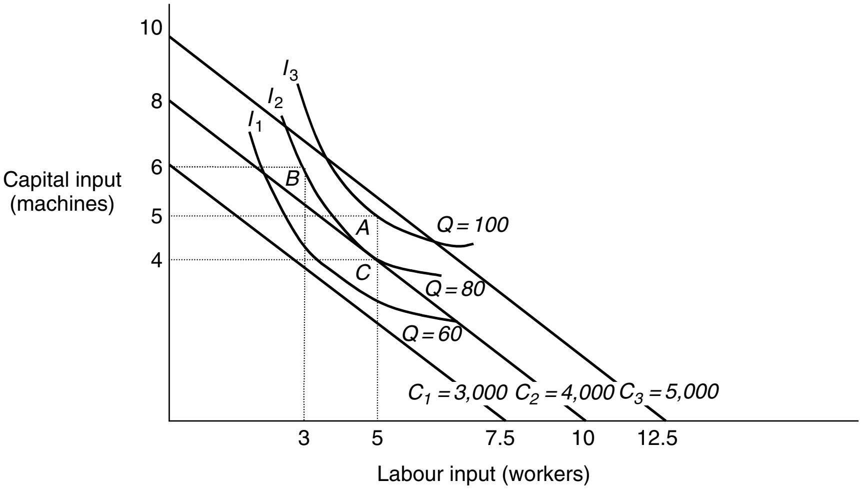

Figure 5.7. Viking Shoes: determining the optimal combination of inputs.

minimum cost to produce an output of 80 units is £4,000, shown by point *C*, and that the input combination required is five workers and four machines. Other combinations of inputs required to produce the same output would cost more than £4,000; for example, the combination at point *B*, three workers and six machines, costs £4,200.

b. Conditions for cost minimization

The cost minimization problem can be examined in more general terms. We have just seen that the condition for optimality is that the isoquant is tangential to the lowest isocost curve. Thus we can equate the slopes of the two curves. The slope of the isoquant is given by the marginal rate of technical substitution, which we have also seen to be given by the ratios of the marginal products, MP_L/MP_k. The isocost line has the equation:

$$C = P_L L + P_K K \tag{5.21}$$

where P_L and P_K represent the prices of labour and capital. The slope of this line, in absolute terms, is given by the ratio of the input prices, P_L / P_K.
Thus:

$$MP_L/MP_K = P_L/P_K, \text{ or } MP_L/P_L = MP_K/P_K \tag{5.22}$$

This means that ***a firm should produce using the combination of inputs such that the ratio of the marginal product of each input to its own price is equal across the last units employed of all inputs***. This principle can be generalized to apply to any number of inputs. It is analogous to the principle in consumer theory that a consumer should spend so that the marginal utility of the last unit of money spent on each product is the same. This was expressed mathematically in (3.14) as:

$$MU_X/P_X = MU_Y/P_Y$$

c. Dual nature of the optimization problem

It has already been indicated that the optimization problem in production theory is in many ways the mirror image of the optimization problem in consumer theory. However, in saying this we are assuming that the nature of the firm's situation is that it has a given target output which it is trying to produce at minimum cost. This is not always the situation. For example, in the public sector the budget may be the given factor and the objective may be to produce the highest output with that given level of budget. This is an output-maximization problem rather than a cost-minimization problem and it exactly parallels the situation of utility maximization in consumer theory. The optimal combination of inputs is again given by the point where the isocost line (in this case a fixed single line) is tangential to the highest isoquant (in this case a variable line). Thus the condition expressed in (5.21) still applies.

d. Changes in input prices

The levels of input prices determine the position and slope of the isocost curves. If the relative prices of the inputs change this will affect the slope of the curves, which we have seen is given by P_L/P_K. If, for example, labour becomes more expensive relative to capital the slope of the isocost curves will become steeper. This will result in the point of tangency moving along the relevant isoquant, upwards and to the left, and a higher level of cost, assuming a given target output. Not surprisingly, less of the more expensive input is used than before, and more of the input that is now relatively cheaper. The situation is illustrated in Figure 5.8. In this example it is assumed that the labour input increases in price from £400 to £500 per week. The isocost curve $C' = 4{,}000$ shows the effect of the price increase and the fact that the output of 80 units can no longer be achieved at the cost of £4,000. To attain this output, assuming economic efficiency, now involves a cost outlay of about £4,400.

There are again obvious parallels in consumer theory, corresponding to the situation where product prices change. In that case it was seen that rational consumers should respond to the situation by buying less of the more expensive product. The main difference is that, because of the dual nature of the situation, consumers are assumed to have a fixed budget line; therefore when a product price rises they are forced onto a lower indifference curve.

e. Expansion paths

Another application of this type of analysis is to consider what happens when the firm's target output increases, or to express the situation in terms of its dual, when the firm's budget increases. As the firm attains higher and higher

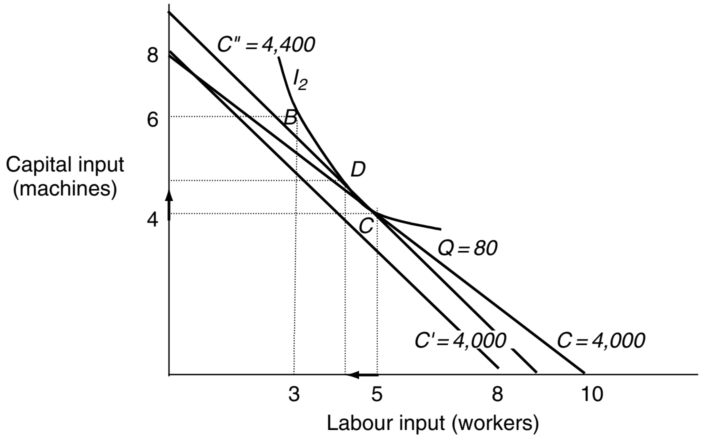

Figure 5.8. Viking Shoes: effects of changes in input prices.

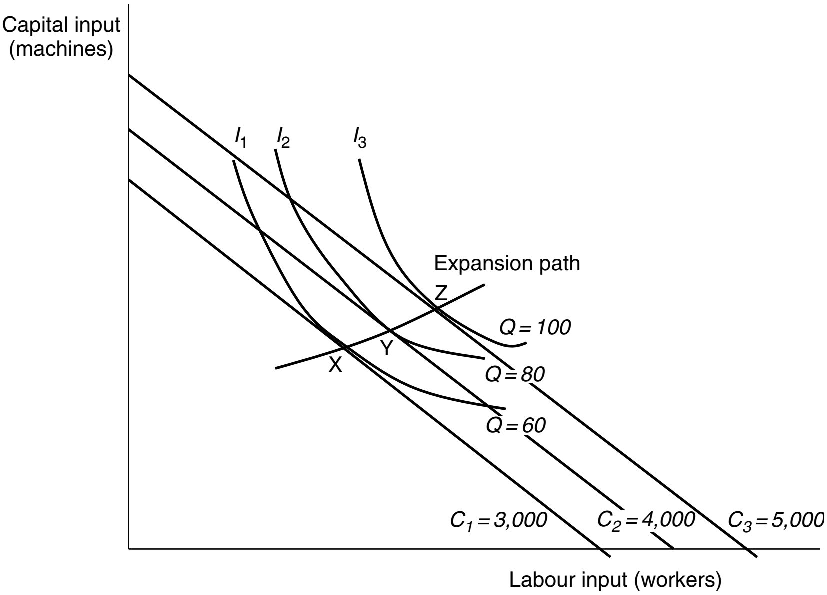

Figure 5.9. Viking Shoes: derivation of expansion path.

output levels the optimal combinations of inputs involved will trace an **expansion path**. This is illustrated in Figure 5.9. The expansion path goes through all the points of tangency, *X*, *Y* and *Z*. This path can be used to determine the long-run relationships between costs and output that are examined in the next chapter. However, the graph in Figure 5.9 assumes that the prices of the inputs remain constant, or at least that their ratio remains constant, which as we shall see is not very realistic.

5.5 A problem-solving approach

It is possible to identify three main management principles that emerge from the preceding discussion of production theory. These are all key points in terms of decision-making.

5.5.1 *Planning*

It can be seen from Table 5.3 that in the short run the range of output for Viking Shoes associated with stage II of the production function is from 27 to 66 units per week. Under most circumstances Viking's optimal operating output should be in this range. If we make the additional assumptions regarding the price of output and the prices of inputs in subsection 5.3.6 we can conclude that optimal output is 59 units; however, it must be remembered that this output is only optimal given the choice of scale by the firm. The implication as far as planning is concerned is that the firm must ensure that it is using the best scale in order to maximize profit. For example, it may be that at the price charged customers might want to buy less than 27 units or more than 66 units, forcing the firm to operate in stage I or stage III. In this situation the firm's scale would be too large or too small respectively. This aspect of planning, **capacity planning**, means that the firm must be able to have accurate forecasts of demand, and communicate the relevant information to its marketing and production departments. These two departments need to communicate with each other, so that sales forecasts by marketing people can be met by the relevant production capacity. Likewise, information relating to production constraints needs to be communicated to the marketing department, so they do not 'oversell' the product.

5.5.2 *Marginal analysis*

This is particularly relevant to optimization problems in neoclassical economic theory. We have now seen various applications of it, in Chapters 2 and 3 and now here in production theory. It will again be important in Chapters 6, 8, 9 and 10. The essential principles are the same in each case. Marginal benefits (profits, utility, product or revenue) tend to be large at low levels of operation (operation can be measured in terms of input, output or expenditure). The operation should be increased until these benefits become equal to the marginal costs. Further expansion is wasteful or non-optimal. The following solved problem will illustrate the application of marginal analysis to production theory.

SP 5.1 Short-run production function

A firm has the following short-run production function:

$$Q = 150L + 18L^2 - 1.5L^3$$

where Q = quantity of output per week
L = number of workers employed.

a. When does the law of diminishing returns take effect?
b. Calculate the range of values for labour over which stages I, II and III occur.
c. Assume that each worker is paid £15 per hour for a 40-hour week, and that the output is priced at £5. How many workers should the firm employ?

Solution

a. $$MP = dQ/dL = 150 + 36L - 4.5L^2$$

MP is at a maximum when diminishing returns occur, therefore we have to differentiate the expression for MP to find the relevant value of L. This is the first-order necessary condition for a maximum.

$$d(MP)/dL = 36 - 9L = 0$$

$$L = 4$$

In order to confirm that this gives a maximum value of MP rather than a minimum we have to consider the second-order condition. This means examining the sign of the second derivative; since this is negative we do indeed have a maximum value of MP when $L = 4$.

b. Stage II begins where AP is at a maximum.

$$AP = 150 + 18L - 1.5L^2$$
$$d(AP)/dL = 18 - 3L = 0$$
$$L = 6$$

Again we can confirm from the second derivative that this gives a maximum.

Stage III begins where $MP = 150 + 36L - 4.5L^2 = 0$

$$L = \frac{-36 \pm \sqrt{\{36^2 - 4(-4.5)(150)\}}}{2(-4.5)} = \frac{-36 \pm 63.21}{-9}$$

$$L = 11.02 \text{ or } 11$$

c. $MRP = MP \times P = (150 + 36L - 4.5L^2)5$ in £ per week
$MFC = 15 \times 40 = 600$

Setting $MRP = MFC$:

$$750 + 180L - 22.5L^2 = 600$$
$$22.5L^2 - 180L - 150 = 0$$
$$L = \frac{180 \pm \sqrt{\{180^2 - 4(22.5)(-150)\}}}{45} = 8.76 \text{ or } 9$$

5.5.3 *Evaluating trade-offs*

We have seen right from the beginning of this book that the concept of opportunity cost is of paramount importance in all areas of economics. In production theory the concept applies to trade-offs between inputs. The most obvious example is the trade-off between capital and labour, as has been seen in the example of Viking Shoes. In this situation the trade-off only applies in long-run time frames. However, many other trade-offs can exist, some of them in the short run.

1 *Labour–labour.* There are different types of labour, with different skills, and these can often be substituted for each other. In a marketing department, for example, salespeople are substitutes for administrative workers to some extent. The managerial problem is to determine the optimal mix of different personnel.

2 *Labour–raw materials.* A good example of this trade-off is in the restaurant business.[7] Many customers may notice that generous portions of condiments and sauces are offered, which may seem wasteful. However, if smaller portions were served, customers would make greater demands of waiters, and the additional labour-time involved might easily offset the savings in terms of raw material costs.

3 *Materials–materials.* Many materials are substitutes for each other. In many cases, substitution of one material for another may affect the quality of the final product, but even if there is no significant difference here, there may well be implications in production. For example, more cars these days are being made out of composite materials. This does affect quality of output in terms of durability, weight and other characteristics; however, there are other implications of substituting composites for metal. Amounts of materials needed and prices are different; the relevant processing of these materials is also different, e.g. moulds are used instead of panel-pressing machinery.

4 *Capital–capital.* Many machines can be substituted for each other, even if their functions are quite different. A manager may have to allocate the

departmental capital budget between photocopy machines, PCs, fax machines and even coffee machines. All contribute directly or indirectly to the total output of the department so again the optimal mix between them must be found.

Many of these trade-offs are relevant in Case Study 5.3 on the National Health Service. The following solved problem also illustrates the situation.

SP 5.2 Optimal combination of inputs

A bottling plant employs three different types of labour: unskilled manual workers, technicians and supervisors. It has estimated that the marginal product of the last manual worker is 200 units per week, the marginal product of the last technician is 275 units per week and the marginal product of the last supervisor is 300 units per week. The workers earn £300, £400 and £500 per week respectively.

a. Is the firm using the optimal combination of inputs?
b. If not, advise the firm on how to reallocate its resources.

Solution

a. The optimal combination is achieved when the marginal product of each type of worker as a ratio of the price of labour is equal, i.e.:

$$\frac{MP_m}{P_m} = \frac{MP_t}{P_t} = \frac{MP_s}{P_s} \tag{5.23}$$

$$\frac{MP_m}{P_m} = 200/300 = 0.67$$

$$\frac{MP_t}{P_t} = 275/400 = 0.6875$$

$$\frac{MP_s}{P_s} = 300/500 = 0.6$$

This combination of inputs is therefore not optimal.

b. It is better to use more of the most productive input, i.e. technicians, and less of the least productive input, i.e. supervisors. By reallocating resources in this way the firm will cause the *MP* of the most productive input to fall and the *MP* of the least productive input to rise, until an optimal point is reached where condition (5.22) is satisfied.

Case study 5.3: Factor substitution in the National Health Service

The National Health Service (NHS) in the UK was founded in 1948 and was the first state-run free health service in the world. It originated at a time of national euphoria following victory in World War II, which generated a sense of confidence and solidarity among politicians and public. In particular it was felt that class distinctions were finally disappearing. The extensive rationing of products, both during and after the war, played a big part. Not only did this result in queuing for goods by rich and poor alike, but it gave the government a sense that state control of distribution was not only possible but in many cases desirable. The basic objective was to provide all people with free medical, dental and nursing care.

It was a highly ambitious scheme that rested on various premises that have since proved flawed. These were:

1 The demand for health care was finite; it was assumed that some given amount of expenditure would satisfy all of the nation's health wants.
2 Health care provision could be made independent of market forces; in particular doctors were not supposed to consider costs in deciding how to treat individual patients.
3 Access to health care could be made equal to all; this means that there would be no preferential treatment according to type of customer, in particular according to their location.

The flaws became more obvious as time went by, and were aggravated by the fact that the system was based on the old pre-war infrastructure in terms of facilities. This meant that the provision was highly fragmented, with a large number of small hospitals and other medical centres. The first flaw became apparent very quickly: in its first nine months of operation the NHS overshot its budget by nearly 40 per cent as patients flocked to see their doctors for treatment. Initially it was believed that this high demand was just a backlog that would soon be cleared, but events proved otherwise. Webster,[8] the official historian of the NHS, argues that the government must have had little idea of the 'momentous scale of the financial commitments' which they had made. Since its foundation, spending on the NHS has increased more than fivefold, yet it has still not kept pace with the increase in demand. This increase in demand has occurred because of new technology, an ageing population and rising expectations. At present it is difficult to see a limit on spending; total spending, public and private, on healthcare in the USA is three times as much per person as in the UK.

However, when it comes to performance compared with other countries the UK does not fare that badly. In spite of far larger spending in the USA, some of the basic measures of a country's health, such as life expectancy and infant mortality, are broadly similar in the two countries. The United States performs better in certain specific areas, for example in survival rates in intensive-care units and after cancer diagnosis, but even these statistics are questionable. It may merely be that cancer is diagnosed at an earlier stage of the disease in the USA rather than that people live longer with the disease.

Performance can also be measured subjectively by examining surveys of public satisfaction with the country's health service. A 1996 OECD study of public opinion across the European Union found that the more of its income that a country spends per person on health, the more content they are about the health service. This showed that, although the British are less satisfied with their health service than citizens of other countries are with theirs, after allowing for the amount of spending per head the British are actually more satisfied than the norm.[9] Italy, for example, spends more per head, yet the public satisfaction rating is far lower.

There are a number of issues that currently face the NHS. The most basic one concerns the location of decision-making. This is an aspect of government policy which is discussed in Chapter 12, and largely relates to normative aspects, though there are some important economic implications in terms of resource allocation. The other issues again have both positive and normative aspects. The use of private-sector providers and charges for services are important issues, again examined in Chapter 12. In terms of spending, once it is recognized that resources are limited, there is the macro decision regarding how much the state should be spending on healthcare in total. Then there is the micro question of where and how this money should be spent, and this issue essentially concerns factor substitution and opportunity cost. A number of trade-offs are relevant here, and some examples are discussed in the following paragraphs.

1. *Beds versus equipment*. Treatments are much more capital-intensive than they used to be in past decades, owing to improved technology. This has the effect of reducing hospital-stay times, and 60 per cent of patients are now in and out of hospital in less than a day[10] compared with weeks or months previously. This can reduce the need for beds compared with equipment.

2. *Drugs versus hospitals*. Health authorities may be under pressure to provide expensive drugs, for example beta interferon for the treatment of multiple sclerosis. This forces unpleasant choices. Morgan, chief executive of the East and North Devon Health Authority, has stated 'It will be interferon or keeping a community hospital, I can't reconcile the two.'[11]

3. *Administrators versus medical staff*. In recent years the NHS has employed more and more administrators, whilst there has been a chronic shortage of doctors and nurses. This was partly related to the aim of the Conservatives when they were in office to establish an internal market (discussed in more detail in Chapter 12). The health secretary, Milburn, was trying to reverse this trend; in a 'top-to-toe revolution' Milburn appeared to want a new modernization board of doctors and nurses to replace the existing board of civil servants. The NHS's chief executive, Langlands, resigned. In the hospitals also there were more administrators, and these took over much of the decision-making previously done by doctors regarding types of treatment. This became necessary because of the clash between scientific advance, increasing costs and budgetary constraints. It became increasingly obvious that rationing had to take place. Related to this issue, nurses were also having to do a lot more administrative work which could be performed by clerical workers. This happened for the same basic reason as before: more information needed to be collected from patients in order to determine the type of treatment.

4. *Hospital versus hospital*. Because of the piecemeal structure that the NHS inherited it has tended to provide healthcare in an inefficient way. Hospitals and other facilities are not only old and in need of repair, but in many cases small, separated geographically, and duplicating facilities. Division of labour is often non-optimal. In Birmingham, for example, there is an accident and emergency unit at Selly Oak Hospital, whereas the brain and heart specialists who might need to perform urgent operations on those involved in car crashes or suffering heart attacks are at the neighbouring Queen Elizabeth Hospital. Thus the issue often arises whether it is preferable to concentrate facilities and staff by building a new and larger hospital to replace a number of older facilities.

5. *Area versus area*. At present there is much variation in the services provided by different local health authorities. For example, some restrict, or do not provide, procedures such as *in vitro* fertilization, cosmetic surgery and renal dialysis. This has led to the description 'postcode prescribing'. Much of this has to do with the differences in budgets relative to demand in different areas, and is another example of the greater visibility of rationing.

Questions

1 Illustrate the trade-off between administrators and medical staff using an isoquant/isocost graph. Explain the economic principles involved in obtaining an optimal situation. How would this situation be affected by an increase in the pay of doctors and nurses?

2 What problems might be encountered in determining this solution in practical terms?

3 Illustrate the hospital-versus-hospital trade-off using an isoquant/isocost graph and explaining the economic principles involved in obtaining an optimal situation. In what important respects does this issue differ from the issue in the previous question?

Summary

1 A production function shows the maximum amounts of output that can be produced from a set of inputs.

2 All points on a production function involve technical efficiency, but only one represents economic efficiency, given prices for the inputs.

3 The functional form of a production function is important because it gives information about the marginal products of the inputs, output elasticities and returns to scale.
4 An isoquant shows different combinations of inputs that can produce the same technically efficient level of output.
5 The marginal rate of technical substitution (*MRTS*) of *X* for *Y* shows the amount of one input *Y* that must be substituted for another *X* in order to produce the same output. It is given by the ratio MP_X / MP_Y and graphically by the slope of the isoquant.
6 Returns to scale describe how a proportionate increase in all inputs affects output; they can be increasing, constant or decreasing.
7 The optimal combination of inputs in the long run is achieved when the marginal product of each input as a ratio of its price is equal.
8 In the short run, production is subject to increasing and then diminishing returns.
9 The optimal level of use of the variable factor in the short run is given by the condition $MRP = MFC$.
10 There are three main applications of production theory in terms of managerial decision-making: capacity planning, marginal analysis and evaluation of trade-offs.

Review questions

1 Give examples of daily activities where the law of diminishing returns applies.
2 Explain the difference between technical and economic efficiency.
3 What is meant by the three stages of production in the short run?
4 Explain the shapes of the total product, marginal product and average product curves for a Cobb–Douglas production function in the short run.
5 Figure 5.10 shows an isoquant and an isocost curve.
Show the effects of the following changes:

a. The price of *L* rises.
b. The prices of *L* and *K* rise by the same proportion.

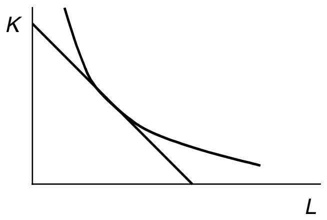

Figure 5.10. Isoquant and isocost curves.