4

PRODUCTION, PRODUCTION FUNCTIONS AND COST CURVES

In the previous chapter the theoretical principles of consumer behaviour were analysed and the demand curve was derived. In this chapter we will start looking at the other side of the market, namely the supply side. The supply side also has a theoretical foundation which is referred to as *production theory*. A number of concepts used in demand theory are applied to analyse production theory.

Production is the process whereby existing factors of production (ie natural resources, labour, capital and entrepreneurship) are used to produce goods and services. Goods, or products, are tangible items such as ballpoint pens, cars or bread; whereas services are intangible, such as the services that a doctor, hairdresser or transport company provides. In this book we will concentrate on the production of tangible items.

The ability and willingness of firms to supply a product on the market depend inter alia on the costs of production. The manufacture of any product involves the use of factors of production which in turn have cost implications. The quantity of a good that a firm would be prepared to supply on the market depends on the price and productivity of the factors of production used, on the one hand, and, on the other hand, on the price that the product will fetch on the market. In this chapter attention will be paid to the general characteristics of production costs. In later chapters product prices will also be considered and the supply decisions of producers will be explained in the light of costs and prices. The price that the producer will fetch for his product is therefore not dealt with in this chapter.

Once you have studied this chapter you should be able to

- describe the various cost and profit concepts
- explain the difference between the short run and the long run
- understand the short-run production function and present it graphically
- use the total product curve to derive the marginal and average product
- understand the long-run production function and the way in which isoquants are used to derive it
- grasp how the different cost curves are derived
- understand the relationship between the different cost curves

THE COST CONCEPT

Opportunity cost

When an economist wishes to determine production costs implicit and explicit costs are both taken into account. What does this entail? *Explicit costs* are the actual expenditures incurred by the firm in order to buy or hire the inputs required in the production process. *Implicit costs,* on the other hand, refer to the value of inputs owned by the firm itself which are also used in the production process. The value of these self-employed inputs must be calculated in terms of what they could have earned in the best alternative use elsewhere.

Implicit costs include the maximum wage or salary which the entrepreneur could have earned in the same position elsewhere (eg as a manager of a similar firm), plus the highest return that the firm could obtain from investing its capital elsewhere and renting out its land and other inputs. The inputs which the firm owns and uses in its own production process cannot be regarded as being free, even if their use is not accompanied by actual (explicit) expenditure.

As has already been mentioned, the costs of using any input by a firm are equal to the highest earnings which the input could be earning elsewhere (outside the firm). This applies to inputs hired by the firm as well as inputs owned by the firm which are used in the production process. If one of the employees of a firm can earn R60 000 in his best alternative occupation (in another firm), the present employer will be obliged to pay that amount. If the employer pays less, the employee concerned will merely terminate his services and go and work for the other firm. The same applies if the entrepreneur can earn more by managing another firm than he is making in his own business – it would be pointless continuing with his own enterprise. (We ignore the high premium that the entrepreneur could place on being his own boss.) Should a firm use any input itself, the costs thereof have to be taken into account by looking at what the input would earn in its best alternative application (outside the firm).

Opportunity costs

Economists' view of costs, known as *opportunity costs,* therefore include explicit and implicit costs. We can say the following:

economic costs of production = explicit costs + implicit costs
= opportunity costs

An example will help highlight the difference between explicit and implicit costs even more. Suppose that Joe Citizen is the sole proprietor of a small estate agency. He owns the building where his office is housed and uses his own labour and capital to operate the business. Although his business does not incur any explicit rental, interest or labour expenses, by implication rent, interest and a

salary are paid. By using his own building as an office Joe is forgoing the rental income that he would have earned if he had rented the building out to someone else. In the same way, by using his capital and labour in his own business, Joe is forgoing the highest interest and salary which he could have earned by applying these resources elsewhere. Joe Citizen's total opportunity costs therefore amount to all the explicit costs (electricity, water, property tax, travel expenses, etc) which he may have *plus* the implicit costs (forgone rent, interest and salary).

It is important to remember that when costs are referred to in this book, the economist's view of costs is assumed. Under certain circumstances these costs may differ markedly from the costs which would be calculated by an accountant.

Normal profit, economic profit and accounting profit

Included in implicit costs is a portion which economists call normal profit. This is the minimum earnings that would be necessary to prevent an entrepreneur from applying his talents and factors of production elsewhere. Stated differently: normal profit is the minimum profit that the owner of a firm would accept in order to become involved in a specific activity and to stay with it. If a firm is making normal profit, it means that the entrepreneur is making exactly as much as he would be making in the best alternative occupation with his own self-employed resources elsewhere. If this minimum profit (or normal profit) is not reached, the entrepreneur will withdraw his factors of production and apply them elsewhere, where the prospects look better. Or, alternatively, the individual will stop being an entrepreneur and go and earn a salary.

Accountants have a somewhat different view of costs, which means that their view of profit also differs from that of economists. For the sake of comprehensiveness the following cases need to be differentiated:

Normal profit

- *Normal profit* is the minimum earnings that will prevent an entrepreneur from employing his factors of production elsewhere. It will therefore be exactly equal to the highest net income which the firm's own self-employed factors of production could earn elsewhere.

Economic profit

- *Economic profit* (also known as excess profit) is the difference between the total income derived from the sale of a firm's output and the opportunity costs (ie explicit and implicit costs, where the latter includes normal profit). If a firm is making economic profit it means that the firm is earning more than it could be earning elsewhere. It is therefore the profit over and above normal profit.

Accounting profit

- *Total or accounting profit* is the difference between a firm's total income from the sale of its product and its explicit costs.

Because of accountants' narrower view of costs, accounting profit is also higher than economic profit. The following should clarify this:

Economic profit = total income – total costs (explicit and implicit)

Accounting profit = total income – total explicit costs

In conclusion: if an economist says that a firm is making a *normal profit*, it means that the firm is covering its economic costs (explicit and implicit costs) and that it is making a profit large enough to retain its factors of production in the present enterprise. If a firm is making more than this, it is earning an *economic profit* (also called an excess profit). A firm can of course also be in a situation where the economic costs surpass the total income – then the firm is making an *economic loss*.

Short and long run

The costs which a firm incurs in the production process will depend on the quantities and combinations of the various factors of production being used. The quantities of some of the factors of production can be changed quickly, whereas others need more time. For example, the capacity of a plant, that is to say the size of the factory and the machinery and equipment that it has at its disposal, can only be adapted over a long period of time. In some industries it can even take years to change the capacity. Differences in the amount of time needed to make such changes have resulted in economists differentiating between the short run and the long run.

Short run: plant size remains unchanged

The short run refers to a time period that is so short that the firm cannot change the size of its production plant, but it can adapt the utilisation of the production plant. Although the firm's capacity remains unchanged over the short run, production can be altered by using more labour and raw materials or less. The existing capacity can therefore be used more intensively over the short run or less.

Long run: plant size changes

From the viewpoint of existing firms the long run refers to a time period that is so long that firms are able to change the quantities of *all their factors of production* – this means that the *capacity* of the production plant can also be changed. From the

viewpoint of the industry the long run refers to a time period whereby new firms can come into operation and enter or leave the industry. If SA Breweries employees 100 extra workers, this will be a short-run adjustment. If the same SA Breweries builds a new factory plant, or expands the existing plant by building on or installing new equipment, this will be a long-run adjustment.

It is important to note that the difference between the short run and the long run is more a conceptual difference than a reference to a specific time period. A small firm that manufactures T-shirts can increase its productive capacity within a few days by buying a few new tables and sewing machines. On the other hand, it can take an enterprise like Iscor years to build a plant. The long run is a few days for the T-shirt manufacturer, whereas for Iscor it is a few years.

THE SHORT-RUN PRODUCTION FUNCTION

An example

How do the quantities that a firm produces change as more and more factors of production are used? The answer lies in the production function, which shows the relationship between the quantity of inputs and the outputs that are obtained from the inputs. The production function can be depicted by means of a table, an algebraic equation or a graph.

To illustrate we will look at the most simple case, namely the short-run production function, where it is assumed that at least one factor of production remains unchanged. We assume that we are concerned with the production of maize, where there is only one variable factor of production (labour) which is combined with a fixed quantity of land (say one hectare) in order to produce maize. We also assume that the state of technology remains unchanged. More and more labourers, all of the same quality, are combined with the fixed quantity of land and the maximum total production (TP) per time period (eg bags of maize per month) is recorded. This information appears in Columns (1) to (3) of Table 4-1. The average product (AP) appears in Column (4) which is the total product divided by the number of labourers; the marginal product (MP), which shows the change in TP if labour increases by one unit, appears in Column (5). The same information can also be depicted graphically, as is done in Figure 4-1.

If total product (or output) is represented by TP and L represents labour units, the production function can be written as

$$TP = f(L) \qquad (1)$$

which merely means that the total product is a function of labour inputs.

Table 4–1

Total, average and marginal product

(one variable factor of production)

(1) Land	(2) Labour	(3) TP	(4) AP_L	(5) MP_L
1	0	0	0	–
1	1	3	3	3
1	2	8	4	5
1	3	12	4	4
1	4	15	3,8	3
1	5	17	3,4	2
1	6	17	2,8	0
1	7	16	2,3	–1
1	8	13	1,6	–3

Law of diminishing returns

A closer observation of Table 4–1 and Figure 4–1 shows that total, average and marginal product initially increase, reach a maximum and then decrease. This phenomenon is known as the *law of diminishing returns*. Formally this law can be stated as follows: *if the quantity of a variable factor of production used in the production process increases while the other factors of production remain constant, levels of production will be reached where first the marginal product, then the average product and finally the total product will begin to decrease over time.*

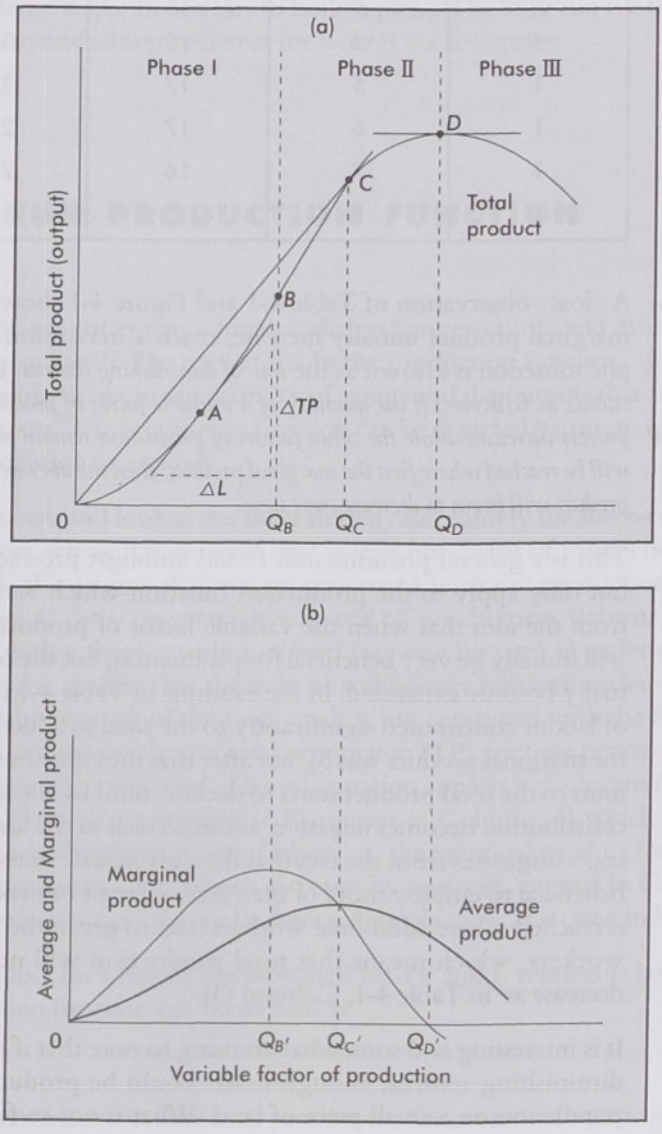
This is a general phenomenon found amongst production functions and does not only apply to the production function which was used here. It originates from the idea that when the variable factor of production (labour) increases, it will initially be very beneficial for production, but the original benefit will eventually become exhausted. In the example in Table 4–1 the use of the second unit of labour contributed significantly to the total product (Column (5) shows that the marginal product was 5), but after that the contribution by additional labour units to the total product starts to decline, until by the seventh unit of labour the contribution becomes negative, as can be seen in the last column. This phenomenon originates from the fact that for a given state of technology it is not always beneficial to employ more of the variable factor of production (labour). A stage is reached where additional workers start to get in the way of and irritate other workers, which means that total production will not increase. It may even decrease as in Table 4–1, Column (3).

It is interesting and somewhat amusing to note that if it were not for the law of diminishing returns, enough maize could be produced for the entire world population on a small piece of land. What is not so funny is that a tragic mis-

Figure 4–1

The relationship between total, average and marginal product

The total product curve appears in the top panel. The point of inflection is at point B where the marginal product changes from increasing to decreasing. At point C the marginal product is equal to the average product, as is shown on the bottom panel at $Q_{C'}$. Beyond point C the average product begins to decrease. At point D the total product reaches a maximum; the marginal product is equal to zero at point D and thereafter becomes negative. Phases I, II and III of the production process are shown in the top panel. Phase II is the economic phase of production.



conception as recently as 40 years ago led to the death of millions of people. Read about this in Box 4–1.

Box 4–1

MAO ZEDONG'S MISCONCEPTION

Mao Zedong was elected president of the People's Republic of China in 1949. Shortly thereafter he formulated an eight-point plan to reform agriculture in his country. The most important part of his plan was to plough deep, plant close together and exterminate pests. Mao conceived that it was logical that if 100 rice plants per square metre produced 100 kilograms of rice, then 10 000 plants (the variable factor of production) would produce 10 000 kilograms on the same piece of land. Naturally, the grain died and people starved on a massive scale.

Until as recently as 1980, after many Chinese had died of starvation, numerous small farmers continued to overcultivate – probably as a result of the continuous propaganda from the authorities. Photographs of grain growing so high and so lush that children could sit on top of it often appeared in the press. A photographer later revealed that the children were actually sitting on a bench.

It is obvious that Mao Zedong and his advisors did not know much about agriculture and they most probably had never heard of the law of diminishing returns.

Source: Mao het 30m Chinese laat sterf, *Insig*, March 1997.

Graphic presentation of the short-run production function

Figure 4–1(a) is a graphic presentation of the law of diminishing returns, whereas Figure 4–1(b) is useful for explaining the relationship between total, average and marginal product. The graph is not an exact representation of the figures that appear in Table 4–1 but, what is more important, it reflects the typical course of the curves. By drawing the curves it is assumed that fractions of labour units can be employed, consequently the curves run smoothly. Note that the total product goes through three phases: initially it rises at an increasing rate up to point B, thereafter it continues increasing, but at a decreasing rate up to point D, finally it reaches a maximum at point D, whereafter it begins decreasing. Point B, where the rate of change in the total product curve alters, is known as the point of inflection. As can be seen in the bottom section of Figure 4–1, point B (the point of inflection) can be described as the point where the marginal product changes from increasing to decreasing.

If the total product is known, it is possible to determine the average and marginal products graphically. The technique for this is used repeatedly later in the chapter and must therefore be understood thoroughly.

Marginal product

The *marginal product* (MP) is defined as the change in total product (ΔTP) which occurs if the variable factor of production, labour, increases by one unit (ie $\Delta L = 1$). We can also state it as follows: the marginal product measures the rate of change in the total product (ΔTP) which occurs with each additional labourer (ΔL) – ie $MP_L = \Delta TP/\Delta L$. It is therefore geometrically equal to the slope of the tangent at any point along the total product curve. If at point A in Figure 4-1 the slope of the tangent is $\Delta TP/\Delta L = 3/1 = 3$, then the value of the marginal product at that point is 3. By drawing more tangents, such as the one at point A along the total product curve, the marginal product can be determined, as is done in Column (5) of Table 4-1. (For mathematically minded students, note that the derivative of the function can also be used.)

The three phases through which the total product passes are also reflected by the marginal product. Refer to Figure 4-1: where total product increases at an increasing rate (up to the point of inflection), the marginal product also increases, that is to say, the extra labourers are making increasing contributions to the total product. Where the total product is increasing at a decreasing rate (after the inflection point) the marginal product is positive, but it is decreasing – every additional labourer is actually making a positive contribution, but is adding less to the total product than the previous labourer did. When the total product reaches a maximum the marginal product is equal to zero, that is to say, the last labourer employed contributes nothing to the total product. When the total product starts to decrease the marginal product is negative – the last labourer employed thus causes the total product to decrease rather than increase.

Average product

The *average product* (AP) is defined as the total product (TP) divided by the number of units of labour (L) used in the production process, that is to say $AP_L = TP/L$. The average product is determined geometrically by calculating the slope of a ray from the origin to a point on the total product curve. If at point C in Figure 4-1 a total of 8 bags of maize are produced ($CQ_C = 8$) and 2 units of labour are used in the process ($0Q_C = 2$), then the slope of the ray from the origin to point C will be equal to $CQ_C/0Q_C$, or $8/2 = 4$; this will reflect the value of the average product at point C. Once again points like point C, with rays from the origin, would put us in a position to calculate the average product in Column (4) in Table 4-1.

The average product reflects the same 'increasing-maximum-decreasing' relationship between variable labour inputs and total output as the marginal product. A definite technical relationship exists between the marginal product and the average product (see Fig 4-1(b)): where the marginal product is greater than the average product, the latter increases; where the marginal product is less

than the average product, the latter decreases. Hence it follows that the marginal product curve intersects the average product curve at its maximum. This relationship is a mathematical one. In terms of our production example it can be stated as follows: as long as the contribution by each additional worker to the total product is greater than the average product of all the workers who are already employed, then the average product will increase. However, if an additional worker's contribution to the total product is less than the average product at that stage, that worker will cause the average product to decrease.

The law of diminishing marginal output is reflected by the shape of all three curves, in other words by the total product, the average product and the marginal product. Economists, however, are most concerned with the marginal product. In Figure 4-1 the three stages of production, which have already been referred to, are therefore differentiated, namely the stages of increasing, decreasing and negative marginal product (or output) – they are reflected as Phases I, II and III. If we look at Columns (2) and (5) of Table 4-1, we will see increasing marginal output for the first two workers, decreasing marginal output for workers 3 to 6 and negative marginal output for workers 7 and 8. A producer will not want to be in Phase III, because there additional labourers (and therefore costs) lead to a decrease in total product (because marginal product is negative). Up to Phase II additional labour leads to an increase in the total product (because the marginal product is positive) – a producer would therefore prefer to be in Phase II, but will not want to move on to Phase III from there.

THE LONG-RUN PRODUCTION FUNCTION

In the previous section the short-run production function was investigated in which one factor of production remained constant and the other one was variable. In this section the situation is observed where a firm uses two factors of production, namely labour (L) and capital (K) and both factors of production are variable. As long as all factors of production are variable, we are concerned with the long-run production function, which is written as follows:

$$Q = f(L,K) \qquad (2)$$

where Q represents the quantity produced.

The technique used in the previous section to analyse the production function with one variable factor of production is somewhat clumsy if there are two variable inputs. We therefore use a new technique which makes use of isoquants. The isoquant concept corresponds with that of indifference curves in many respects.

Definition of isoquants

An isoquant shows the different combinations of labour (L) and capital (K) with which a firm can produce a specific output. A higher isoquant (further from the origin) represents a higher output and a lower isoquant a smaller output.

Properties of isoquants

As has already been mentioned, isoquants have many similarities to indifference curves, which were dealt with in Chapter 3. There are, however, also important differences between isoquants and indifference curves which must be taken into account.

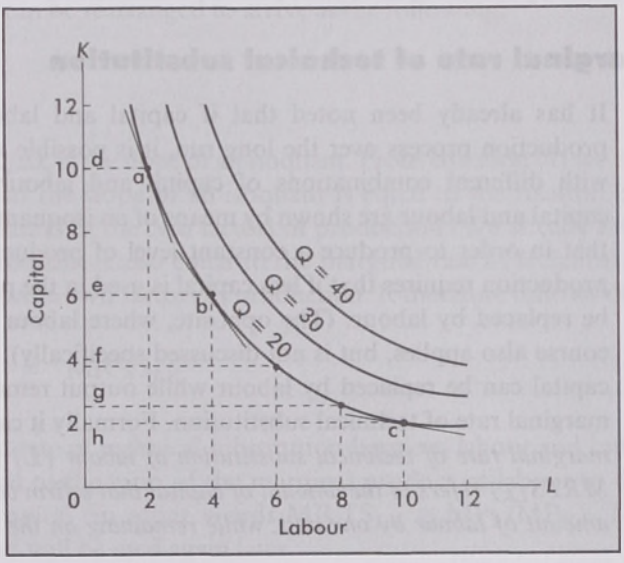
As is already clear from the definition, isoquants show all the combinations of two inputs (factors of production) which, given the production function, produce a specific output. Three isoquants which show how capital and labour are combined to produce three different quantities of output appear in Figure 4-2 – these are only three of an infinite number of isoquants which we could have drawn. The isoquants in the diagram are depicted as Q = 20, Q = 30 and Q = 40 respectively, which refer to the output Q which each one represents. For example, the isoquant Q = 20 shows all the possible combinations of capital and labour which, according to the production function, will produce 20 units of output. In so doing 10 units of capital combined with 2 units of labour (point a in Fig 4-2), or 6 units of capital combined with 4 units of labour (point b in Fig 4-2) will produce an output of 20 units each time. An important difference between indifference curves and isoquants now comes to the fore – the production which an isoquant indicates can be quantified, that is to say it can be depicted in quantifiable units, whereas the level of satisfaction depicted by an indifference curve cannot be quantified.

Isoquants (just like indifference curves) can *never intersect or touch* each other. If two isoquants were to intersect each other, then the same quantity of capital and labour would produce two different levels of output at the point of intersection. This is impossible, seeing that the production function shows the *maximum* output that can be achieved with any combination of inputs. The isoquant which represents the smaller output therefore cannot exist at the intersection. The same argument holds for a point of tangency.

In order to produce a constant level of output, the technical aspect of production requires that if less capital is used in the production process, it must be replaced with labour. This requirement of *technical substitution* has resulted in isoquants having a negative slope, that is to say the relevant part of an isoquant slopes downwards from left to right. The reason for this will become clear in the next paragraph. (Why we talk about the *relevant* part will become clear under the heading 'Ridge lines' on page 100.)

Figure 4–2
An isoquant map

Over the long run both capital and labour can vary in the production process. An isoquant shows different combinations of capital and labour with which a specific output will be produced. Isoquants further away from the origin represent a larger output.



Isoquants are *convex* when viewed from the origin of a graph, that is to say the curves form a 'bulge' which points towards the origin. This property arises because although technical substitution can take place between capital and labour, the substitution is not perfect. This means that if the firm reduces the amount of capital by equal amounts in the production process, the amount of labour that would be required to keep output constant would become more and more. This is also illustrated in Figure 4–2. If the amount of capital required to produce 20 units is decreased from 10 to 6 units, the amount of labour must increase from 2 to 4 units (compare points a and b). A reduction of 4 units of capital thus means that 2 more units of labour need to be used to continue producing an output of 20. If we now move from point b to point c in the diagram, we see that a further decrease of 4 units of capital will result in labour having to be increased by 6 units to keep output constant. To move from point b to point c the firm therefore has to substitute more labour per unit of capital than when moving from point a to b. (The same principle applied to indifference curves.)

Another (reverse) method of explaining the same result is as follows: for a constant level of output the quantity of capital which is replaced by an additional amount of labour will decrease as more and more substitution takes place. This is obvious because in our example the relation of the change in labour to the change in capital

($\Delta L/\Delta K$) was originally 2:4 (ie 1:2) and thereafter changed to 6:4 (ie 1:2/3). Between a and b 1 unit of labour could replace 2 units of capital, but between b and c 1 unit of labour could only replace 2/3 units of capital. This property of the production process is called the principle of the *diminishing marginal rate of technical substitution*. It will be discussed more formally in the next section.

The marginal rate of technical substitution

It has already been noted that if capital and labour are combined in the production process over the long run, it is possible to obtain the same output with different combinations of capital and labour. These combinations of capital and labour are shown by means of an isoquant. We have also mentioned that in order to produce a constant level of production, the technical side of production requires that if less capital is used in the production process, it must be replaced by labour. (The opposite, where labour is replaced by capital, of course also applies, but is not discussed specifically). The relationship whereby capital can be replaced by labour while output remains constant is called the marginal rate of technical substitution. Formally it can be stated as follows: the *marginal rate of technical substitution of labour (L) for capital (K) (or rather $MRTS_{LK}$) refers to the amount of capital that a firm can give up by increasing the amount of labour by one unit, while remaining on the same isoquant.*

The value of the *marginal rate of technical substitution* of labour for capital at any point along an isoquant is also shown by the absolute value of the *slope* ($\Delta K/\Delta L$) at that point (the negative sign in front of the slope is ignored here and in the rest of the chapter). If we observe points a, b, and c in Figure 4–2, we notice that the absolute value of the slope of the isoquant Q = 20 decreases as we move from point a to b to c. This tendency will continue as we move downwards along an isoquant and we therefore say that the *marginal rate of technical substitution of labour for capital ($MRTS_{LK}$)* decreases. As has already been explained, one unit of labour will be able to replace fewer and fewer units of capital as there is a movement down along the isoquant. (Also study Box 4–2.) It can also be seen in the fact that in Figure 4–2 for *equal increases* in labour on the horizontal axis, the amount of capital replaced on the vertical axis gets *smaller* – compare the decreasing distances between points d, e, f, g and h.

The marginal rate of technical substitution can also be expressed in terms of the marginal product (ie marginal output). The total production, when moving along an isoquant, can only remain unchanged because the production that the producer is 'giving up' by using less capital ($\Delta K \times MP_K$) is exactly cancelled out by the production that the producer 'gains' by using more labour ($\Delta L \times MP_L$). This means that the following condition must be upheld (once more we ignore the signs):

$$\Delta K \times MP_K = \Delta L \times MP_L \quad (3)$$

where ΔK = the change in the quantity of capital
MP_K = the marginal product of capital
ΔL = the change in the quantity of labour
MP_L = the marginal product of labour

Equation (3) can be rearranged to arrive at the following:

$$\frac{\Delta K}{\Delta L} = \frac{MP_L}{MP_K} \quad (4)$$

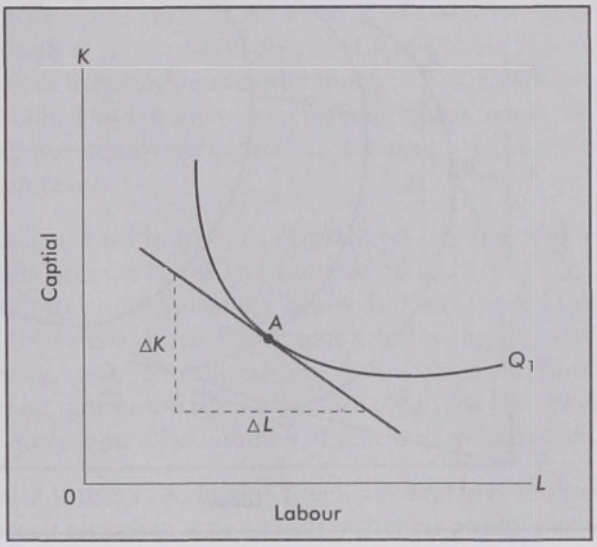
The term $\Delta K/\Delta L$ is the slope of an isoquant. From equation (4) we can therefore determine that the slope of an isoquant is equal to the relation between the marginal products of the two factors of production. We already know that the slope of an isoquant is also equal to the marginal rate of technical substitution (MRTS) between two factors of production. It therefore follows that:

$$\frac{\Delta K}{\Delta L} = \frac{MP_L}{MP_K} = MRTS_{LK} \quad (5)$$

The marginal rate of technical substitution between labour and capital is therefore also equal to the ratio of the marginal product of labour to the marginal product of capital (in other words $MRTS_{LK} = MP_L/MP_K$). This result is important and will be used again later.

Box 4–2

SUMMARY: THE SLOPE OF AN ISOQUANT



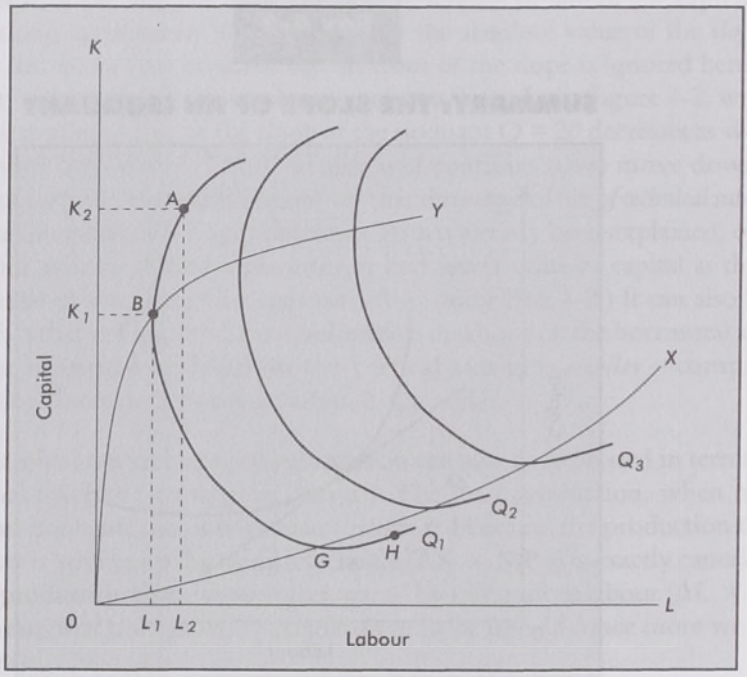
The slope of an isoquant shows the ratio $\Delta K/\Delta L$, which is the ratio whereby capital is replaced by labour while output remains constant. It is known as the marginal rate of technical substitution. At point A the following applies: $MRTS_{LK} = \Delta K/\Delta L = MP_l/MP_k$.

Ridge lines

An isoquant can also have sections with a positive slope, in other words that rise from left to right, like the parts AB and GH in Figure 4-3. A firm, however, will not produce at these points on an isoquant, because the same output can be produced by using less of both labour and capital. For example, at point A in Figure 4-3 more labour and more capital is being used than at point B – the same output can therefore be produced more cheaply at point B than at point A. The same argument holds for points G and H in Figure 4-3. Lines through such points which separate the relevant sections (ie with negative slopes) from the irrelevant sections (ie with positive slopes) of isoquants are called *ridge lines*. In Figure 4-3 0Y and 0X are examples of ridge lines. The area between the ridge lines is called the *technically efficient area* and is the part of the isoquants along which the firm would want to produce.

Figure 4-3
Ridge lines

0Y and 0X are ridge lines. The area between the ridge lines is the technically efficient area where a firm will produce.



Economies and diseconomies of scale

Economics of scale

Diseconomies of scale

The isoquant concept makes it possible to explain at this stage an economic concept which comes under discussion later in the chapter. In practice it is a general phenomenon that the unit costs of production will decrease as a firm expands by building a bigger plant and increasing production volumes. This is known as internal *economies of scale* and occurs because mass production gives rise to the possibility of lower unit costs of production as a result of factors such as specialisation and the use of better technology. A firm can, however, get too big and reach a stage where additional expansion will lead to greater unit costs – this is known as *diseconomies of scale*.

The concepts of economies and diseconomies of scale are illustrated with the aid of isoquants in Figure 4–4. In each of the diagrams a ray with a 45° angle is drawn. The axes have the same scale, and if there is a movement along the ray the factors of production will increase proportionately. To illustrate we allow the quantity of capital and labour to increase by 100% in each case (from 10 to 20 units) and see what happens to production volumes as a result. The isoquants which appear in each of the diagrams present exactly the same production volumes (100, 200 and 300 units).

Constant returns to scale. In diagram (a) the distance between the three isoquants is constant. If the quantity of capital and labour increases by 100% in this instance, then the output also increases by 100% (from 100 to 200 units). The increase in the output is therefore exactly the same as the increase in the factors of production (in percentage terms). This is known as constant returns to scale and in this case the firm experiences neither economies nor diseconomies of scale.

Increasing returns to scale. In diagram (b) the successive distances between the isoquants become smaller. An increase of 100% in the quantity of capital and labour leads to an increase in output of 200% (from 100 to 300 units) in this case. The output therefore increases by more, in percentage terms, than the factors of production. This is known as increasing returns to scale and in such a case a firm will experience *economies of scale* if it expands. Economies of scale give rise to lower unit costs.

Decreasing returns to scale. In diagram (c) the successive distances between the isoquants become larger. An increase of 100% in the quantity of capital and labour in this case results in a below 100% increase in the quantity of output (from 100 to 180 units). The output therefore increases by a smaller amount, in percentage terms, that the factors of production do. This situation is known as decreasing returns to scale and in such a case the firm experiences *diseconomies of scale* if it expands. Diseconomies of scale lead to higher unit costs.

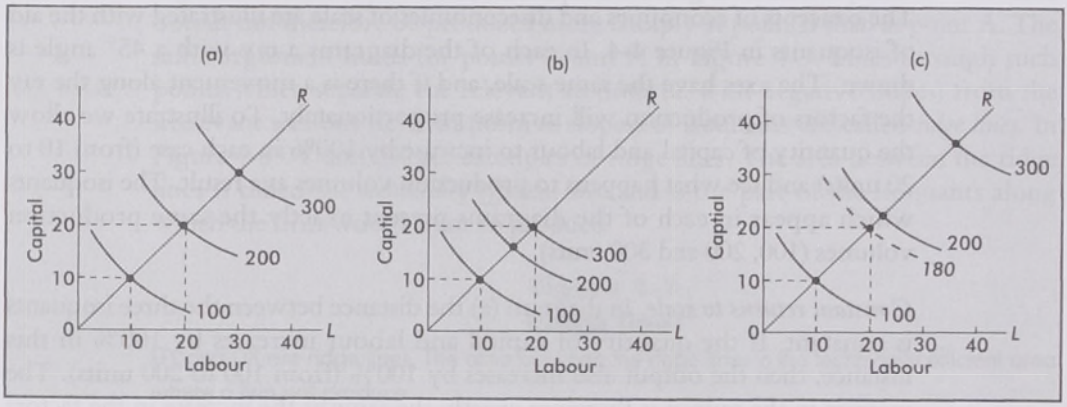
It is possible that a production function could give rise to increasing, decreasing or constant returns to scale. In reality all three possibilities would probably occur

at different stages of production. Initially, when production volumes are small, it can be expected that a firm which expands will experience increasing returns to scale. Thereafter it will most probably experience constant returns to scale. Finally, with high production volumes, decreasing returns to scale will be experienced. As will be seen later, this phenomenon has an influence on the shape of the long-run cost curves.

Figure 4–4

Constant, increasing and decreasing returns to scale

Figure (a) illustrates constant returns to scale, whereas Figure (b) illustrates increasing and Figure (c) decreasing returns to scale.



OPTIMAL INPUT COMBINATIONS

In a previous section we saw that if two inputs are used in the production process, the firm will produce in the technically efficient area of an isoquant. This information, however, is not sufficient to determine the exact combination of inputs which the firm will employ in order to deliver a given output, the reason being that an infinite number of possible input combinations exist within the technically efficient region of any isoquant. To show exactly how a firm decides on an input combination, production and cost theory must be combined. To do this it is assumed that in order to produce any given output the firm will choose the combination of inputs that will minimise costs of production; such an input combination is known as an *optimal input combination*. Before proceeding any further with the optimal input combination, we must first introduce a new concept, namely isocost curves. This is a concept which concurs with the budget line which we came across in Chapter 3.

Isocost curves

An isocost curve (also referred to as an isocost line) shows the different combinations of labour and capital that a firm can buy, given the amount of money (also

called the total outlay) that the firm has available and the current prices of the factors of production.

Observe a firm that has a total amount of R1 000 to spend on labour and capital and that the price of labour (P_L) and the price of capital (P_K) are both R10 per unit. (We could have used bigger, more realistic amounts, but the principles remain the same.) Table 4-2 shows some ways in which the firm could have spent R1 000.

Table 4-2

Combinations of capital and labour

(R1 000 is available. P_K = R10 per unit, P_L = R10 per unit)

Combination	Capital (K) (units)	Labour (L) (units)	Total costs (R)
a	100	0	1 000
b	80	20	1 000
c	60	40	1 000
d	40	60	1 000
e	20	80	1 000
f	0	100	1 000

Isocost curve

The combinations in Table 4-2 (as well as the combinations in between, such as 70 units of capital and 30 units of labour) are illustrated graphically in Figure 4-5 by means of the straight line which is drawn from point a to point f. At a the firm buys only capital and no labour, whereas at point f only labour and no capital is purchased. This line is known as an *isocost curve* (or equal cost curve) because it shows the combinations of capital and labour that the firm can acquire for a given outlay.

In reality the intercepts on the axes are all that are necessary to draw the isocost curve, that is to say the maximum number of units of each factor of production that the firm can afford if the available amount of money is spent in its entirety on that factor of production. If the firm spends its entire outlay (TO) on capital, it can buy TO/P_K units of capital – in the example that is R1 000/R10 = 100 units. In the same manner the firm, if it only buys labour, can purchase TO/P_L units, which also works out as R1 000/R10 = 100. By joining these two points on the axes we get the isocost curve of the firm. As has already been said, the firm can purchase any combination of labour and capital along the isocost curve. Every isocost curve represents a given outlay, in other words, what is available to spend on labour and capital. In the example the total outlay is R1 000. If it were to increase to, say, R2 000 while the prices of labour and capital remained the same, a new isocost curve, further from the origin, would be obtained.

The *slope of the isocost curve* is as follows:

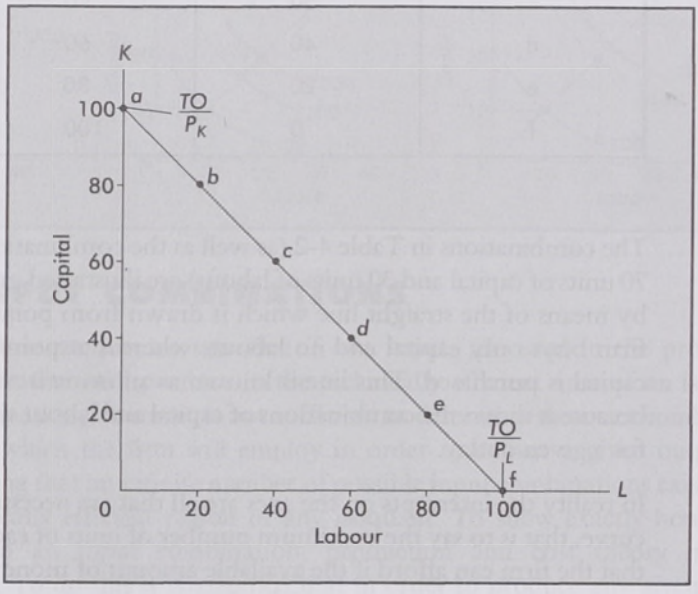
$$\frac{TO/P_K}{TO/P_L} = \frac{TO}{P_K} \cdot \frac{P_L}{TO} = \frac{P_L}{P_K} \quad (6)$$

The slope of the isocost curve is equal to the relation of the price of labour to the price of capital, that is to say P_L/P_K. In the example which we are using $P_L = P_K = R10$ and $TO = R1\,000$. This means that the isocost curve in Figure 4-5 has a slope of 1. The slope is calculated as follows: $P_L/P_K = R10/R10 = 1$.

Figure 4-5

Isocost curves

They show the combinations of capital and labour that a firm can buy for a given outlay. If the intercepts on the axes (TO/P_K and TO/P_L) are calculated, the isocost curves can be drawn. The slope of the isocost curve is P_L/P_K.



Choice of input combination

We can now determine how a firm will choose the optimal input combination, that is the combination of inputs that will minimise production costs, or – which amounts to the same thing – we can look at how a firm chooses its input combination in order to produce the maximum output for a given cost. To do this a diagram which depicts the isocost curve and the isoquants of a producer is used.

Equilibrium

A producer will be in *equilibrium* if he produces a maximum output with a given cost outlay. Another way of putting this is that a producer is in equilibrium if, with a given isocost curve, he reaches the highest possible isoquant. This happens when the isocost curve just touches an isoquant, in other words forms a tangent to the isoquant. This is illustrated in Figure 4–6, where isoquant Q_2 is the highest isoquant which can be reached by the isocost curve concerned – the producer is therefore in equilibrium at point M. Isoquant Q_3 cannot be reached with the given isocost curve and if the firm produces on isoquant Q_1 it will not maximise output.

At the point of tangency M on Figure 4–6 the slope of the isoquant ($MRTS_{LK}$) is equal to the slope of the isocost curve (P_L/P_K). It therefore applies that in equilibrium $MRTS_{LK} = P_L/P_K$. From Equation (5) it is also known that $MRTS_{LK} = MP_L/MP_K$ and it therefore applies that for a *producer to be in equilibrium*

$$\frac{MP_L}{MP_K} = \frac{P_L}{P_K} \text{ or } \frac{MP_L}{P_L} = \frac{MP_K}{P_K} \qquad (7)$$

This is an important result – it means that the *producer is in equilibrium* if the marginal product (MP) of the last rand spent on labour is exactly the same as the marginal product (MP) of the last rand spent on capital. If the firm uses more than two factors of production, the same will apply (in the equilibrium situation) for the other factors of production.

If the relationship between the marginal product and the price (eg MP_L/P_L) is not the same for all factors of production, the producer can reach a higher level of production by changing the combination of the factors of production. If the marginal product (or marginal output) which the producer obtains from the last rand spent on labour is greater than the marginal product that he obtains from the last rand spent on capital, he can increase total production by spending more on labour and less on capital. If, on the other hand, the relationship between the marginal product and the price is the same for both factors of production, then production cannot be increased further and production equilibrium (ie maximum production) is reached. That is the best that the producer, given his cost constraint, can do.

The expansion path

Earlier in the chapter it was said that the long run of the firm is under consideration if all the factors of production, capital included, are variable. An implication of this is that in order to produce a certain output the firm can expand and can build the plant size that will minimise costs. If it is assumed that the firm uses two factors of production, labour and capital, and that their prices remain constant if the firm expands its production activities, we can determine how the firm will act if it increases production over the long run.

In Figure 4–7 the production function of the firm is presented by means of four isoquants Q_1, Q_2, Q_3 and Q_4. Four isocost curves, C_1 to C_4, are also drawn on the diagram. These isocost curves are equally spaced and all have the same slope because the relationship of the price of labour to the price of capital (ie P_L/P_K) remains constant. The further away the isocost curve lies from the origin, the larger the output it represents. The isocost curves are drawn in such a way that they form tangents to the isoquants at points H, I, J and K – each of these points also represents a point of equilibrium for the firm, in other words the maximum output that can be produced with the costs concerned. By joining these equilibrium points a curve is obtained which shows how the input combinations which minimise costs change for increasing levels of production. Stated differently: this curve, which is called the *expansion path* of the firm, shows the *lowest possible input costs (combinations of K and L) at which different levels of production can be produced.* The benefit of this information will become clear in the next section. (The similarity between the expansion path and the income-consumption curve which was explained in Chapter 3 should already be clear at this stage.)

Expansion path

Figure 4–6

Production equilibrium

The firm is in equilibrium at point M on isoquant Q_2. That is the highest isoquant that can be reached by the isocost curve. At M the following applies: $MRTS_{LK} = P_L/P_K = MP_L/MP_K$.

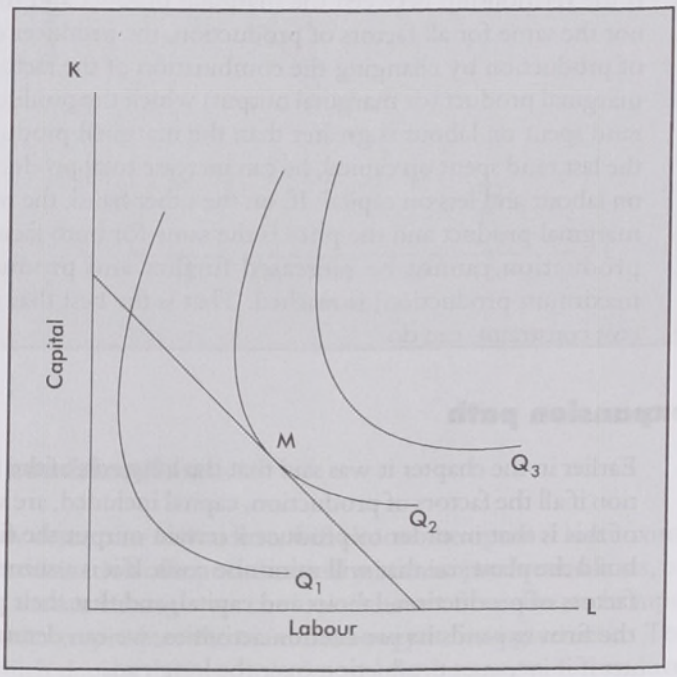
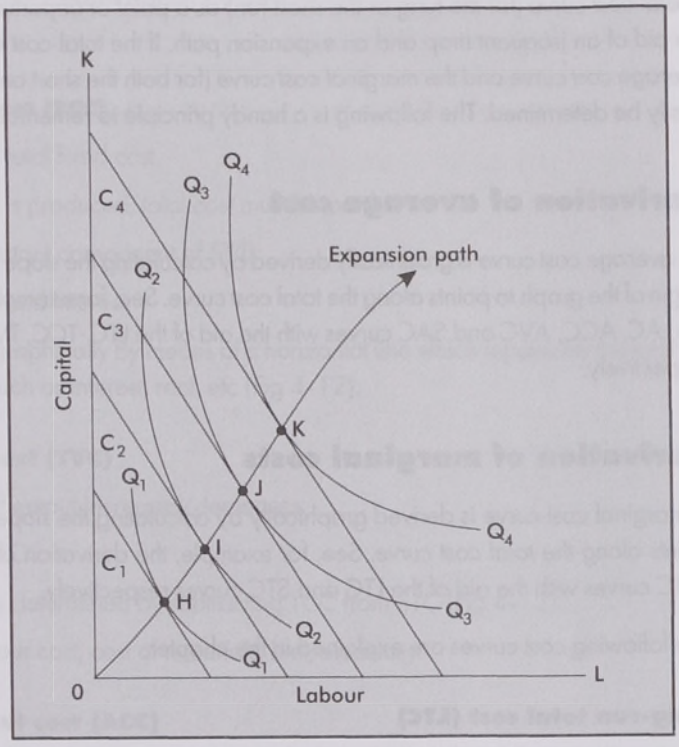


Figure 4–7

The expansion path of the firm

The curve shows the lowest possible input costs (combinations of K and L) at which different output levels can be produced.



Box 4–3

COST CURVES: A SUMMARY

In the rest of the chapter the various cost curves of the firm are derived. The following is a summary of all the cost curves. You will probably only understand it fully once you have studied the rest of the chapter, but it will be of benefit to you to work through this section beforehand. Once you have studied the chapter it will serve as a good summary.

No fewer than ten different cost curves are derived below. For many a student this is an intimidating thought. Note, however, that as soon as the total cost curves (for both the long and the short run) are determined, the average and marginal cost curves easily result from them.That leaves only four cost curves that need to be studied. Of the latter, the constant cost curve should pose the fewest difficulties; to derive the average constant

cost curve from it is easy. All that remains to be studied is the variable cost (total and average). Seen in this light, the determination of cost curves is not such a big task.

The derivation of the cost curves of a firm can be carried out graphically by constructing a *total cost curve* (for the long or the short run) as a point of departure; this is done with the aid of an isoquant map and an expansion path. If the total cost curve is known, the *average cost curve* and the *marginal cost curve* (for both the short and the long run) can easily be determined. The following is a handy principle to remember:

Derivation of average cost

An average cost curve is graphically derived by calculating the *slope of the rays* from the origin of the graph to points along the total cost curve. See, for example, the derivation of the LAC, ACC, AVC and SAC curves with the aid of the LTC, TCC, TVC and STC curves respectively.

Derivation of marginal costs

A marginal cost curve is derived graphically by calculating the *slope of the tangents to points along the total cost curve*. See, for example, the derivation of the LMC and the SMC curves with the aid of the LTC and STC curves respectively.

The following cost curves are explained in the chapter:

Long-run total cost (LTC)

- It is illustrated graphically by using the information about costs and production volumes (output) which is reflected by the expansion path (Figs 4–7 and 4–8(a)).
- Total cost includes explicit cost plus implicit cost (the latter includes normal profit).

Long-run average cost (LAC)

- $LAC = LTC/Q$
- It is graphically determined by calculating the slope of rays that are drawn from the origin to points along the LTC curve (Figs 4–8(a) and 4–8(b)).

Long-run marginal cost (LMC)

- $LMC = \Delta LTC/\Delta Q$
- Definition: the LMC is the addition to long-run total costs (ΔLTC) if one additional unit of output (ΔQ) is produced.
- It is graphically determined by calculating the slope of tangents to points along the LTC (Figs 4–8(a) and 4–8(b)).

Short-run total cost (STC)

- It is illustrated graphically by using the information in terms of the costs and quantities of production reflected by the expansion path (Figs 4–10 and 4–11).
- Short-run total cost consists of total constant cost plus total variable cost: STC = TCC + TVC.

Total constant cost (TCC)

- Also known as total fixed cost.
- Even if nothing is produced, total cost must be paid.
- It forms the constant component of STC.
- Examples: rent and interest.
- It is presented graphically by means of a horizontal line which represents the total of constant cost such as interest, rent, etc (Fig 4–12).

Total variable cost (TVC)

- TVC changes if output increases/decreases.
- TVC = STC – TCC
- It is graphically determined by subtracting TCC from STC (Fig 4–12).
- Examples: labour cost, cost of materials, fuel, electricity.

Average constant cost (ACC)

- ACC = TCC/Q
- ACC decreases as long as output increases.
- It is graphically determined by calculating the slope of the rays which are drawn from the origin to points along the TCC (Figs 4–13(a) and 4–13(b)).

Average variable cost (AVC)

- AVC = TVC/Q
- It is graphically determined by calculating the slopes of the rays that are drawn from the origin to points along the TVC curve (Figs 4–14(a) and 4–14(b)).

Short-run average cost (SAC)

- SAC = STC/Q
- It is graphically determined by calculating the slopes of rays that are drawn from the origin to points along the STC curve (Figs 4–14(a) and 4–14(b)).

Short-run marginal costs (SMC)

- SMC = $\Delta STC/\Delta Q$ or
- SMC = $\Delta TVC/\Delta Q$
- Definition: the SMC is the addition to short-run total costs (ΔSTC) if one additional unit of output (ΔQ) is produced.
- Alternative definition: the SMC is the addition to total variable cost (ΔTVC) if one additional unit of output (ΔQ) is produced.
- The SMC is graphically determined by calculating the slopes of tangents to the STC curve or TVC curve (Figs 4-14(a) and 4-14(b)).

COSTS

The expansion path which was derived in Figure 4-7 reflects information that can be used to derive the cost curves of a firm. If the cost curves of a firm are available, it can be seen how firms decide how much they are going to produce and supply to the market under various market conditions (this is done in later chapters). The determination of the cost curves of a firm is a relatively technical process and is sometimes repetitive in nature. At the end of the chapter practical methods which are used to determine cost curves are also discussed. Before we do this we must study the underlying principles of the various cost curves. This can be done in two ways. First, use can be made of numerical tables which show what the relationship between the various costs is and how they are calculated. Second, the manner in which the various costs are derived and how they are related to one another can be explained graphically. The latter method is followed in the rest of this chapter (it is probably the easier method).

When studying the rest of the chapter it is advisable to continuously refer to the summary in Box 4-3.

Long-run total cost (LTC)

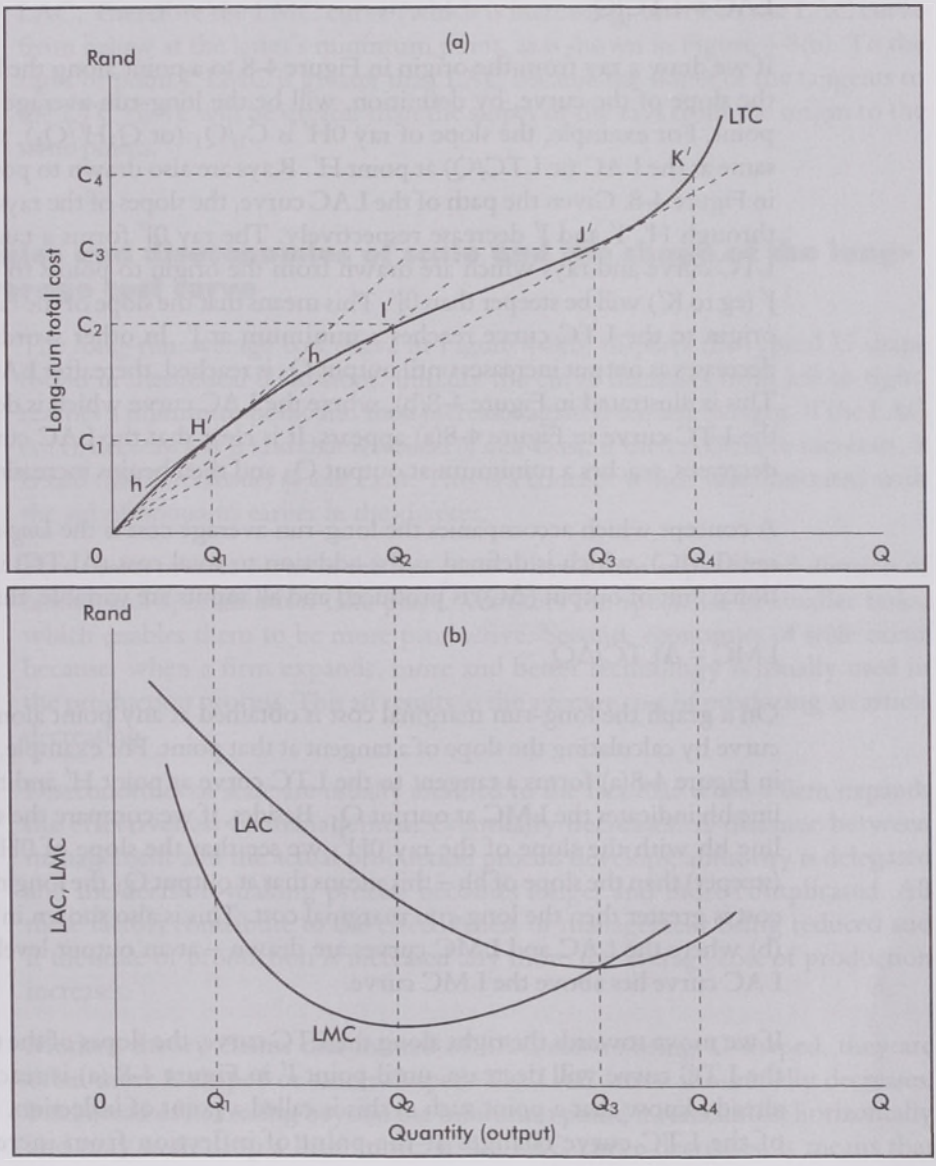
Seeing that the expansion path shows the lowest possible costs at which various output levels can be produced when all factors of production are variable, it can be used to construct the long-run total cost curve (LTC); this is done by plotting the costs and the corresponding quantities of production on a graph. This is done in Figure 4-8, where point H′ provides the same information as point H in Figure 4-7. Point H in Figure 4-7 shows that C_1 is the lowest cost at which Q_1 can be produced. In Figure 4-8 H′ shows exactly the same thing, that is to say Q_1 and C_1 are used as coordinates to plot H′. (In both Fig 4-7 and Fig 4-8 Q_1 for example will indicate 100 000 tins of beer and C_1 a total cost of R250 000.) In the same manner points I′, J′ and K′ in Figure 4-8 are plotted by using the output

and total cost represented by points I, J and K in Figure 4–7 respectively. The curve which is drawn from the origin through points H′, I′, J′ and K′ in Figure 4–8 is the long-run total cost curve (LTC).

Figure 4–8

Derivation of the long-run total cost, long-run marginal cost and long-run average cost

The expansion path is used to calculate the long-run total cost curve. This curve enables us to determine the LMC and LAC.



Long-run average cost (LAC) and long-run marginal cost (LMC)

Once the long-run total cost curve has been determined it can be used to calculate the long-run average cost as well as the long-run marginal cost. The method used to do this is the same as that used earlier in the chapter to derive the marginal product and the average product curve from the total product curve.

The long-run average cost (LAC) is the long-run total cost (LTC) which is necessary to produce a given output divided by that output, that is to say

$$LAC = LTC/Q \tag{8}$$

If we draw a ray from the origin in Figure 4–8 to a point along the LTC curve, the slope of the curve, by definition, will be the long-run average cost at that point. For example, the slope of ray $0H'$ is C_1/Q_1 (or Q_1H'/Q_1), which is the same as the LAC (ie LTC/Q) at point H′. Rays are also drawn to points I′ and J′ in Figure 4–8. Given the path of the LAC curve, the slopes of the rays which pass through H′, I′ and J′ decrease respectively. The ray $0J'$ forms a tangent to the LTC curve and rays which are drawn from the origin to points to the right of J′ (eg to K′) will be steeper than $0J'$. This means that the slope of the rays from the origin to the LTC curve reaches a minimum at J′. In other words, the LAC decreases as output increases until output Q_3 is reached, thereafter LAC increases. This is illustrated in Figure 4–8(b), where the LAC curve which is derived from the LTC curve in Figure 4–8(a) appears. It is clear that the LAC curve initially decreases, reaches a minimum at output Q_3 and then begins increasing.

A concept which accompanies the long-run average cost is the *long-run marginal cost* (LMC), which is defined as the addition to total cost (ΔLTC) if one additional unit of output (ΔQ) is produced and all inputs are variable, that is to say

$$LMC = \Delta LTC/\Delta Q \tag{9}$$

On a graph the long-run marginal cost is obtained at any point along the LTC curve by calculating the slope of a tangent at that point. For example, the line hh in Figure 4–8(a) forms a tangent to the LTC curve at point H′ and the slope of line hh indicates the LMC at output Q_1. Besides, if we compare the slope of the line hh with the slope of the ray $0H'$, we see that the slope of $0H'$ is greater (steeper) than the slope of hh – this means that at output Q_1 the long run average cost is greater then the long-run marginal cost. This is also shown in Figure 4–8 (b) where the LAC and LMC curves are drawn – at an output level of Q_1 the LAC curve lies above the LMC curve.

If we move towards the right along the LTC curve, the slopes of the tangents to the LTC curve will decrease, until point I′ in Figure 4–8 (a) is reached – you already know that a point such as this is called a point of inflection. (The slope of the LTC curve changes at the point of inflection from increasing at a

decreasing rate to increasing at an increasing rate.) After point I′ the slopes of the tangents once more become steeper (not shown in the diagram) and the LMC therefore increases, which results in the LMC curve reaching a minimum at Q_2, that is to say at that output which corresponds with the point of inflection. This is clearly shown in the diagram. Seeing that the slope of ray 0I′ is still greater than the slope of a tangent to the LTC at I′, it follows that at an output level of Q_2 the LAC is still greater than the LMC. At J′, as we have already discussed, the LAC reaches a minimum value. Because the slope of ray 0J′ is the same as the slope of a tangent at point J′ would be, it follows that for an output level of Q_3 LMC = LAC. Therefore the LMC curve (which is increasing) intersects the LAC curve from below at the latter's minimum point, as is shown in Figure 4-8(b). To the right of point J′ LMC is greater than LAC because the slopes of the tangents to the LTC curve will be greater than the slopes of the rays from the origin to the same points.

Economies and diseconomies of scale and the shape of the long-run average cost curve

Economies and diseconomies of scale

The long-run average cost curve in Figure 4-8(b) displays the typical U shape found in theoretical discussions. Initially the curve decreases from left to right, reaches a minimum point and then increases again from left to right. If the LAC curve decreases, it is said that *economies of scale* exist; if the LAC curve increases, it is said that *diseconomies of scale* exist. This is a concept which was illustrated with the aid of isoquants earlier in the chapter.

Economies of scale occur in the first place when a firm expands and division of labour and specialisation take place. Workers can specialise in smaller tasks, which enables them to be more productive. Second, economies of scale occur because, when a firm expands, more and better technology is usually used in the production process. This all results in the average cost of producing an article decreasing.

Diseconomies of scale are usually ascribed to the fact that when a firm expands the effectiveness of management eventually decreases. A distance between management and the actual production process develops, authority is delegated and the decision-making process becomes longer and more complicated. All these factors contribute to the effectiveness of management being reduced and if the scale of production is increased any more the average cost of production increases.

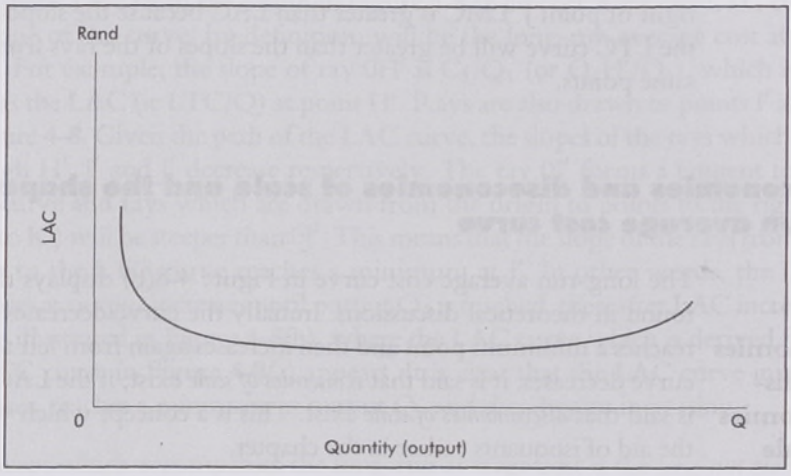
Modern theory claims that instead of LAC curves being U-shaped, they are often more L-shaped or saucer-shaped. The LAC curve also initially decreases, but instead of increasing beyond the minimum point, it runs almost horizontally and may even drop a little more; if the LAC curve decreases, it means that

economies of scale can be reached up to very high production volumes. If the LAC curve does indeed begin to rise at even greater production volumes, it will have a saucer-shaped path. Such a curve is shown in Figure 4–9.

Figure 4–9

Economies of scale and the shape of the LAC curve

Modern theory claims that LAC curves are often L-shaped instead of U-shaped. Sometimes the LAC curve only begins increasing at very high production volumes, in which case it is saucer-shaped.



Short-run total costs (STC)

It has already been shown how the long-run expansion path of the firm is derived with the aid of isoquants and isocost curves and how the long-run expansion path is used to determine the long-run total cost curve. The short-run total cost curve (STC) curve can also be derived with the aid of isoquants and isocost curves.

In Figure 4–10 the production function of a firm is illustrated and the expansion path shows the route along which the firm will expand over the long run. The question now is how can the figure be used to acquire information about costs over the short run? It is already known that over the short run one or more of the factors of production is constant – if we therefore assume labour to be variable and keep capital constant at KK on Figure 4–10, the total costs over the short run at various levels of production can be determined. For example, if the firm wishes to produce the quantity depicted by isoquant Q^*, point B will be the lowest cost at which production can take place. Because B lies on KK (remember

that the latter indicates that capital is constant), point B shows the lowest cost at which $Q^{\star}$ can be produced over the *short run*. The quantity produced and the cost at point B can be used to plot one point along the short-run total cost curve, namely point D in Figure 4–11. (Point B also lies on the long-run expansion path and at the same time provides a point along the long-run cost curve; it can therefore be seen in Figure 4–11 that point D is common to both STC and LTC.)

More points like point B can be determined, which can then be used to plot other points along the STC curve. If the firm for example wants to produce Q_0, it will produce at point A′ in Figure 4–10 in the short run. Over the long run the firm would prefer to produce at point A, but because capital is constant over the short run, the firm must (over the short run) produce at A′ which lies along KK. If we draw an isocost curve through A′, it will show higher costs than the one which is drawn through A. Therefore because A′ lies to the right of the isocost which forms a tangent to Q_0 at A, it means that the total cost to produce Q_0 over the short run is higher than the cost of producing Q_0 over the long run; this is clear in Figure 4–11. In the same way, if the firm wants to produce Q_2 it will produce at point C in Figure 4–11 over the long run, but at point C′ over the short run. Because C′ lies to the right of the isocost which forms a tangent to Q_2 at C, the total cost to produce Q_2 in the short run is higher than the total cost to produce Q_2 over the long run (also see Figure 4–11). If a sufficiently large number of isoquants and isocost curves are drawn in Figure 4–10 and more points like A′, B and C′ are obtained, a continuous STC curve can be drawn, as is done in Figure 4–11.

Just as the long-run total cost curve is used to derive the long-run average and marginal cost, so too the *short-run total cost curve* can be used to derive the *short-run average* and *marginal* cost. This is done later in the chapter.

Total constant cost (TCC) and total variable cost (TVC)

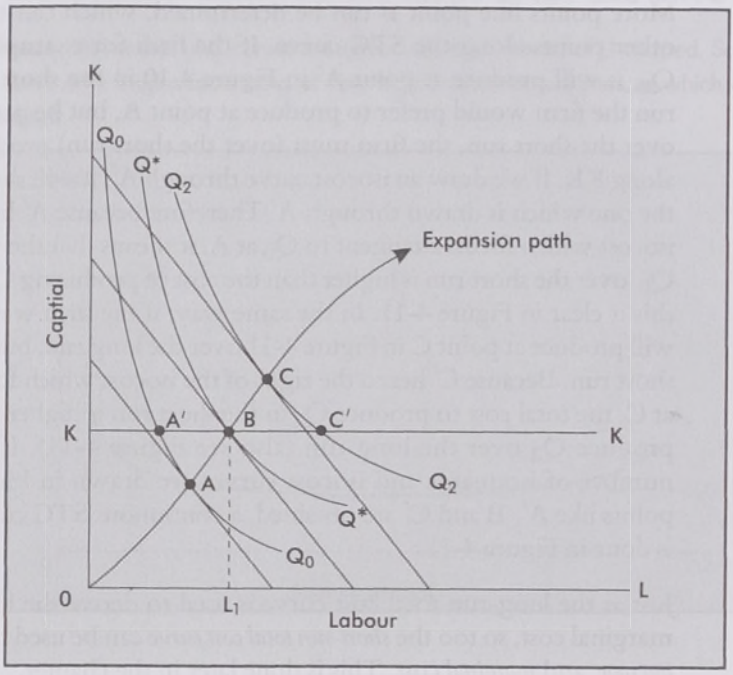
The short-run total cost of a firm can be divided into two parts, namely total constant cost and total variable cost. This can also be written as follows: STC = TCC + TVC. This division is especially important for the management of any enterprise, as will be explained.

Even if a firm produces nothing, there are costs which must be incurred over the short run. This cost is called *total fixed cost or total constant cost* (TCC) – it forms the constant component of STC and is also the reason that the STC curve in Figure 4–11 intersects the vertical axis at a positive value. Total constant cost does not change as the firm's output varies and even if nothing is being produced, expenditures such as rent, interest, depreciation, insurance and salaries for the management must be paid. Because constant cost does not change, the TCC curve (see Fig 4–12) is a horizontal line, irrespective of the output.

Figure 4–10

Short-run production

By keeping capital constant along KK, the short-run total cost can be determined at various output levels. The information can be used to determine the STC curve.



Total variable cost, as the name indicates, is that cost which changes as output varies. Cost items such as labour, cost of materials, fuel, electricity and transport costs are included. If the TCC and the STC curves are known (see the previous section), we are in a position to determine the *total variable cost* (TVC): that is the difference between short-run total cost (STC) and total constant cost (TCC). The TVC is determined graphically by simply subtracting the total constant cost (TCC) from the short-run total cost (STC), as is shown in Figure 4–12.

Differentiating between constant and variable costs is particularly important for the management of any enterprise, the reason being that management can control or vary variable cost over the short run by changing the level of production (the importance of this is discussed again in Chapter 5). On the other hand, constant cost is beyond the control of management over the short run – it must be paid irrespective of the level of production. Constant cost is sometimes referred to as unavoidable cost, fixed cost, overhead cost or indirect cost.

Figure 4–11

Comparison of the long-run total cost curve with the short-run total cost curve

The STC curve lies above the LTC curve. The STC curve does not start at the origin because of the constant cost component which applies over the short run.

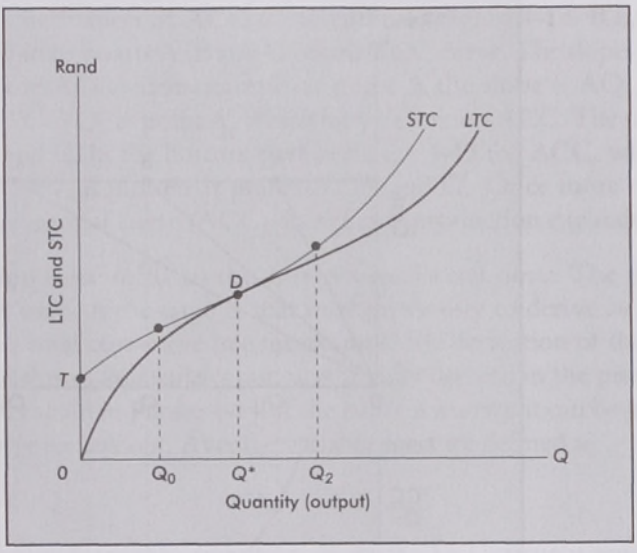


Figure 4–12

The total constant cost curve and the total variable cost curve

TCC does not change if the firm's output varies and the curve therefore runs horizontally. The TVC curve is determined graphically by subtracting TCC from the STC.

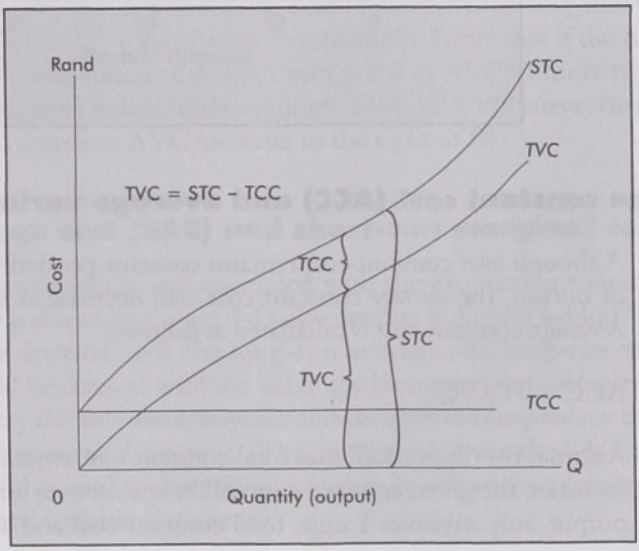
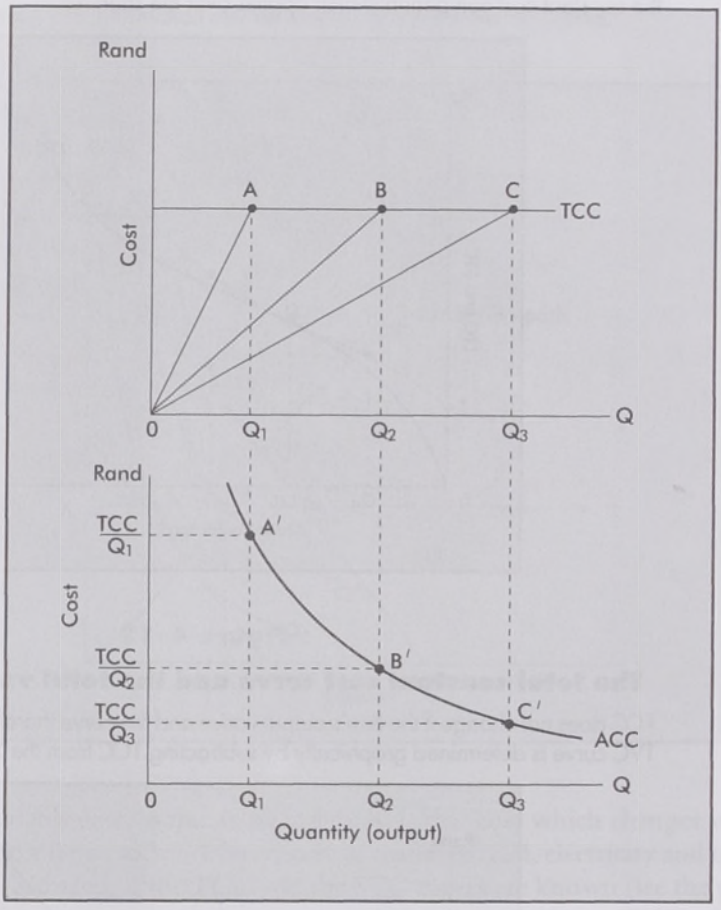


Figure 4–13

The average constant cost curve

The slope of the rays from the origin to points such as A, B and C on the TCC curve shows the average constant cost.



Average constant cost (ACC) and average variable cost (AVC)

Although *total* constant cost remains constant per definition and is independent of output, the *average* constant cost will decrease as long as output increases. Average constant cost is calculated as follows:

$$ACC = TCC/Q \qquad (10)$$

Assume, for illustration, that total constant cost amounts to R1 000. As output increases, the given constant cost will be spread over an increasing output. If the output only involves 1 unit, total constant cost and average constant cost are

both R1 000 (ie ACC = TCC/Q = 1 000/1). If output increases to 2 units, total constant cost is still R1 000, but ACC then becomes R500 (ie ACC = 1 000/2) per unit produced; then R333,33 if R1 000 is spread over 3 production units; R250 if R1 000 is spread over 4 production units, etc. Business people refer to this as the 'spreading of overhead costs'.

The graphic derivation of ACC is presented in Figure 4-13. Rays are drawn from the origin to points A, B and C on the TCC curve. The slopes of these rays amount to the ACC – for example at point A the slope is $AQ_1/0Q_1$ which amounts to TCC/Q_1 at point A, which is by definition ACC. The same happens at points B and C. In the bottom part of Figure 4-13 the ACC, which is calculated in this way, is plotted at points A′, B′ and C′. Once more note that the average constant cost curve (ACC) decreases as production expands.

The attention now shifts to the *average variable cost curve*. The principle for deriving the curve is the same as that used previously to derive an average cost curve from a total cost curve (see for example the derivation of the LAC). The curve which shows *total variable cost*, was already derived in the previous section and it appears again in Figure 4-14. If the curve is known it can be used to calculate the *average variable cost*. Average variable costs are defined as

$$AVC = TVC/Q \tag{11}$$

If rays are drawn from the origin to intersect the TVC curve at points such as A, C and D in Figure 4-14(a), the slopes of the rays will give the AVC at the levels of production that correspond with the points of intersection. For example, the slope of ray 0L at point A is equal to $AQ_1/0Q_1$ which is also equal to TVC/Q_1 and therefore indicates the average variable cost at A, which is also reflected by A′ in Figure 4-14(b). At point C the slope of the ray is smaller than at A and therefore the AVC is lower at C′ than at A′; in the same way the AVC at D′ is lower than it is at C′. It can also be seen from the figure that if the AVC is calculated at D, it will indicate the minimum point of AVC; if more rays are drawn from the origin to points to the right of D on the TVC curve, their slopes will increase and therefore AVC increases to the right of D′.

Short-run average cost (SAC) and short-run marginal cost (SMC)

The STC curve which was associated with the production function in Figure 4-10 and was shown in Figure 4-11 also appears in Figure 4-14(a). Earlier in the chapter we showed how the long-run average and long-run marginal cost curves could be derived with the aid of the long-run total cost curve. In exactly the same way the *short-run average cost* and the *short-run marginal cost* can be derived from the *short-run total cost curve*. (The method to derive the SAC is also exactly the same as that used in the previous section to derive the average variable cost from the total variable cost curve.)

The short-run average cost is defined as

$$SAC = STC/Q \tag{12}$$

To determine the SAC graphically we can once more refer to Figure 4-14(a). Rays are drawn from the origin to points such as B and E on the STC curve; the slopes of these rays are equal to the SAC at the corresponding output levels. This enables us to draw the SAC curve in the bottom part of the figure; the curve reaches a minimum at E′ because the slope of the ray to the STC curve reaches a minimum at E in Figure 4-14 (a).

The short-run marginal costs of a firm can be defined as the increase in short-run total costs (ΔSTC) if one additional unit of output (ΔQ) is produced. Mathematically this can be stated as follows:

$$SMC = \Delta STC/\Delta Q \tag{13}$$

Alternatively short-run marginal costs can be defined as the increase in total variable costs (ΔTVC) if one additional unit of output (ΔQ) is produced, that is to say

$$SMC = \Delta TVC/\Delta Q \tag{14}$$

Both definitions of short-run marginal costs are correct because, as we have already seen, the difference between STC and TVC is a constant, namely total constant cost (TCC). Diagrammatically the paths of the STC and the TVC curves are therefore the same with a constant difference between them – as can be seen in Figure 4-14(a). To determine the SMC at any output, we merely draw a tangent to the STC or the TVC curve at the output level under consideration. The slope of any of the tangents represents the SMC. In Figure 4-14(a) it can be seen that the slopes of the tangents at A and F are the same, giving the SMC at output Q_1 – in the bottom part of the diagram the SMC is represented by A″. In the same way points such as B″ and the rest of the SMC curve can be determined. At B the slope of a tangent to the STC curve (and therefore the value of SMC) reaches a minimum, as is reflected by B″.

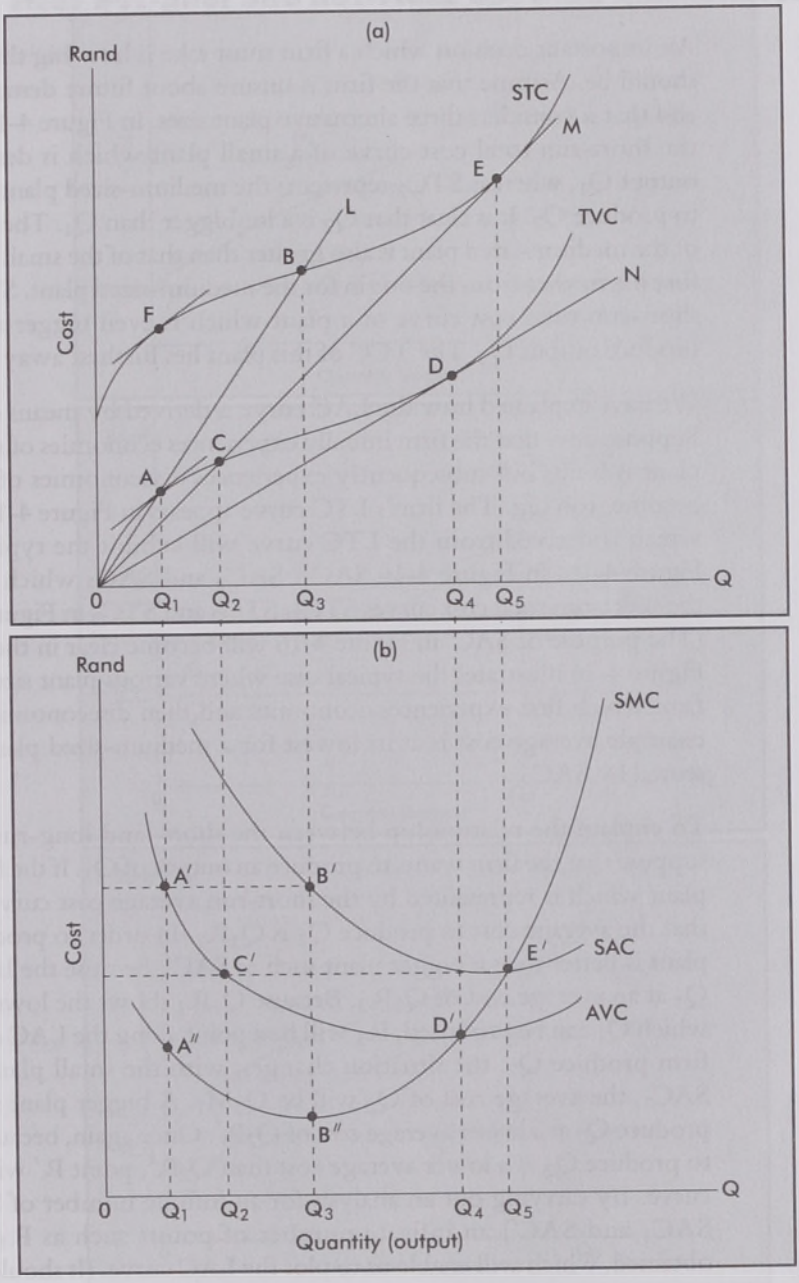
Note that in Figure 4-14(a) the tangents at points D and E are also rays 0N and 0M from the origin. Therefore SMC is equal to SAC at output Q_5 (see point E′) and SMC is equal to AVC at output Q_4 (see point D′). This is not merely a coincidence, but a mathematical fact which results from the path of the STC and TVC curves.

Figure 4-14(b) is an important diagram. It shows all the short-run curves which were derived and their relation to each other. The SAC and the AVC curves are U-shaped. The SAC curve always, at any output level, lies above the AVC curve and reaches a minimum at a higher level of production. The SMC curve is also U-shaped and reaches a minimum value at an output level which

Figure 4–14

The average variable cost curve

The slopes of the rays from the origin to points such as A, C and D on the TVC curve in Figure 4–14(a) show the average variable cost at these points; this is shown at A′, C′ and D′ respectively in Figure 4–14(b). The SAC and SMC curves also appear in Figure 4–14(b).



corresponds with the point of inflection on either the STC or the TVC curve. The SMC curve intersects the AVC and the SAC curves from the bottom at their minimum values.

Relationship between short-run and long-run costs

An important decision which a firm must take is how big the production plant should be. Assume that the firm is unsure about future demand for its product and that it considers three alternative plant sizes. In Figure 4–15 STC_1 represents the short-run total cost curve of a small plant which is designed to produce output Q_1, whereas STC_2 represents the medium-sized plant which is designed to produce Q_2. It is clear that Q_2 is a lot bigger than Q_1. The total constant cost of the medium-sized plant is also greater than that of the small plant, TCC therefore lies further from the origin for the medium-sized plant. STC_3 represents the short-run total cost curve of a plant which is even bigger and is designed to produce output Q_3. The TCC of this plant lies furthest away from the origin.

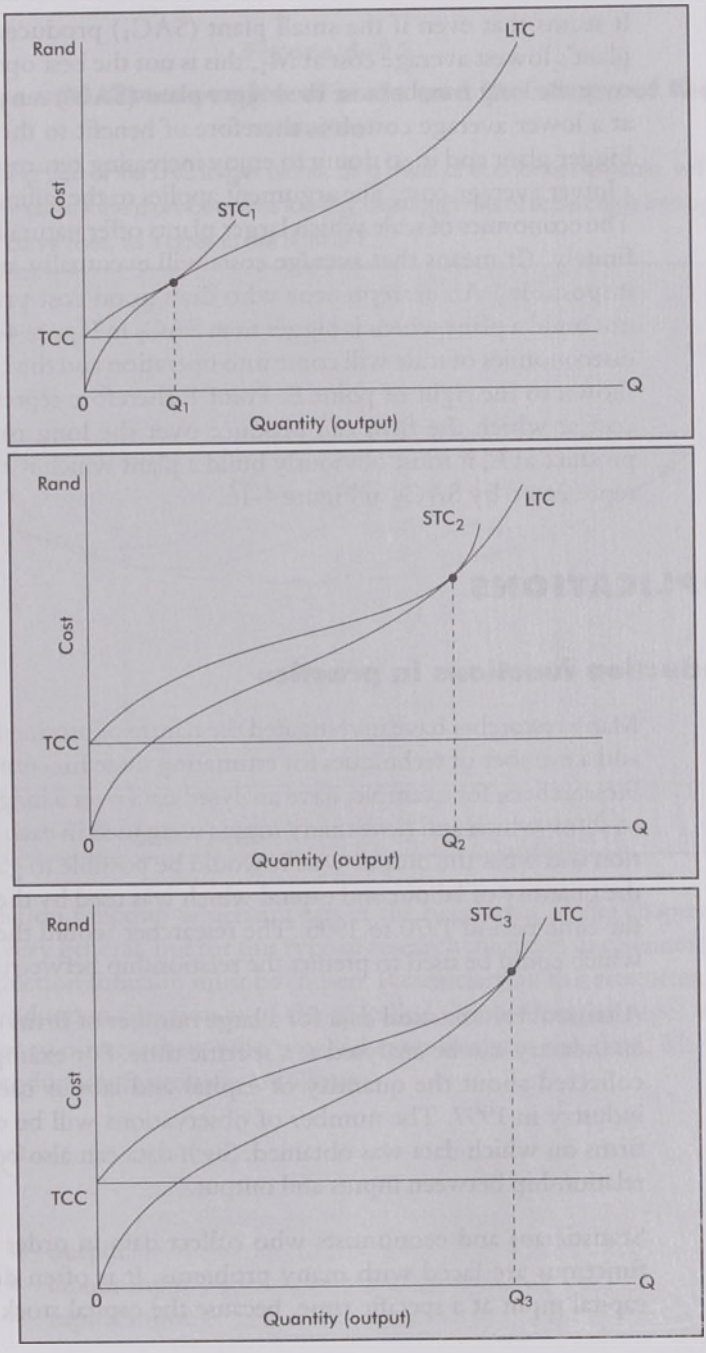
We have explained how the LAC curve is derived by means of the LTC curve. Suppose now that the firm initially experiences economies of scale when a larger plant is built, but subsequently experiences diseconomies of scale as the plant becomes too big. The firm's LTC curve appears in Figure 4–15; the LAC curve which is derived from the LTC curve will exhibit the typical U-shape as in Figure 4–16. In Figure 4–16 SAC_1, SAC_2 and SAC_3 which are derived from the short-run total cost curves STC_1, STC_2 and STC_3 in Figure 4–15 are shown. (The purpose of SAC′ in Figure 4–16 will become clear in the next paragraph.) Figure 4–16 illustrates the typical case where various plant sizes are shown for a firm which first experiences economies and then diseconomies of scale. In this example average cost is at its lowest for a medium-sized plant which is represented by SAC_2.

To explain the relationship between the short- and long-run cost curves, we suppose that the firm wants to produce an output of Q_1. If the firm uses the small plant which is represented by the short-run average cost curve SAC_1, it means that the average cost to produce Q_1 is Q_1R_1. In order to produce Q_1 the small plant is better than a bigger plant such as SAC′, because the latter will produce Q_1 at an average cost of Q_1R_2. Because Q_1R_1 shows the lowest average cost at which Q_1 can be produced, R_1 will be a point along the LAC curve. Should the firm produce Q_2, the situation changes: with the small plant represented by SAC_1, the average cost of Q_2 will be Q_2M_1. A bigger plant such as SAC′ can produce Q_2 at a lower average cost of Q_2R'. Once again, because it is impossible to produce Q_2 at a lower average cost than Q_2R', point R' will lie on the LAC curve. By carrying out an analysis for an infinite number of SAC curves (like SAC_1 and SAC′) an infinite number of points such as R_1 and R' will be obtained, which will enable us to plot the LAC curve. (It should already be clear

Figure 4–15

The long-run total cost curve and the short-run total cost curve

Three different plant sizes are represented by STC_1, STC_2 and STC_3. LTC represents the long-run total cost curve of the firm throughout.



by now that this method also provides an alternative way of deriving the LAC curve. This comes under discussion again later.) At this stage it is easy to see why the LAC curve is also called the envelope curve – it envelops or surrounds the SAC curves in a manner of speaking.

It seems that even if the small plant (SAC_1) produces the quantity Q_2 at the plant's lowest average cost at M_1, this is not the best option for the entrepreneur over the long run, because the bigger plant (SAC′) can produce the quantity Q_2 at a lower average cost. It is therefore of benefit to the entrepreneur to build a bigger plant and in so doing to enjoy increasing returns to scale by producing at a lower average cost. The argument applies to the falling part of the LAC curve. The economies of scale which larger plants offer naturally cannot continue indefinitely. (It means that average costs will eventually amount to zero, which is impossible.) An entrepreneur who does good cost predictions however, will not build a plant which is bigger than SAC_2 in Figure 4-16, because beyond that diseconomies of scale will come into operation and the LAC curve increases, as is shown to the right of point E. Point E therefore represents the lowest average cost at which the firm can produce over the long run. If the firm wants to produce at E, it must obviously build a plant which is as big as the one which is represented by SAC_2 in Figure 4-16.

APPLICATIONS

Production functions in practice

Many researches have investigated the nature of production functions in practice and a number of techniques for estimating these functions have been developed. Researchers, for example, have analysed data over a long period of time in order to show which and how many inputs were used in each period under investigation and what the output was. It would be possible to gather information about the quantity of labour and capital which was used by the clothing industry over the time period 1970 to 1995. The researcher would then have 26 observations which could be used to predict the relationship between inputs and output.

Alternatively statistical data for a large number of firms or for various sectors of an industry can be analysed at a specific time. For example, information can be collected about the quantity of capital and labour used by firms in the tyre industry in 1997. The number of observations will be equal to the number of firms on which data was obtained. Such data can also be used to determine the relationship between inputs and output.

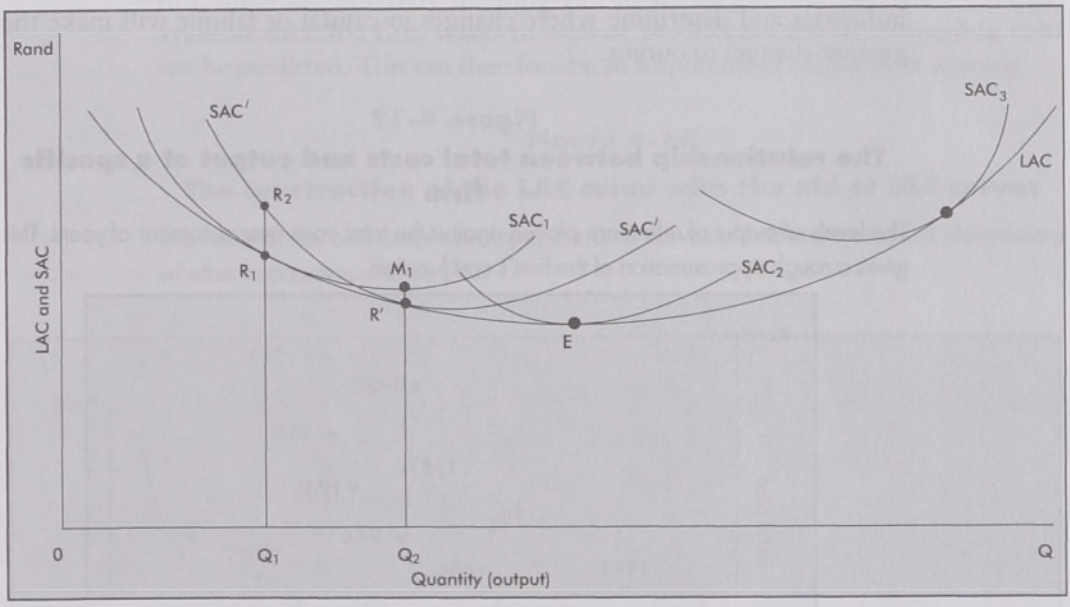
Statisticians and economists who collect data in order to predict production functions are faced with many problems. It is often difficult to measure the capital input at a specific time, because the capital stock consists of equipment

which is not identical, nor is it the same age, nor is it equally productive. Even the labour input is difficult to measure because labour quality in a firm can vary considerably. To discuss all the problems concerning the calculation of production functions can take up volumes, but there is no need to go deeper here.

Figure 4–16

Economies and diseconomies of scale and the shape of the LAC curve

Along the falling part of the LAC larger plants, as a result of economies of scale, will lead to lower average costs. If the plant becomes too big, diseconomies of scale come into operation and the LAC curve rises, as it does to the right of E.



Cobb-Douglas

The production function which appears at the beginning of the chapter (equation (2)) is very general, but for this type of research the exact mathematical form of the production function must be chosen. Researchers in this area often assume that the production function is of the so-called Cobb-Douglas type, which is named after two researchers who worked extensively in this field. The Cobb-Douglas production function is as follows:

$$Q = AL^{\alpha}K^{\beta} \qquad (15)$$

where Q = output
L = labour input
K = capital input

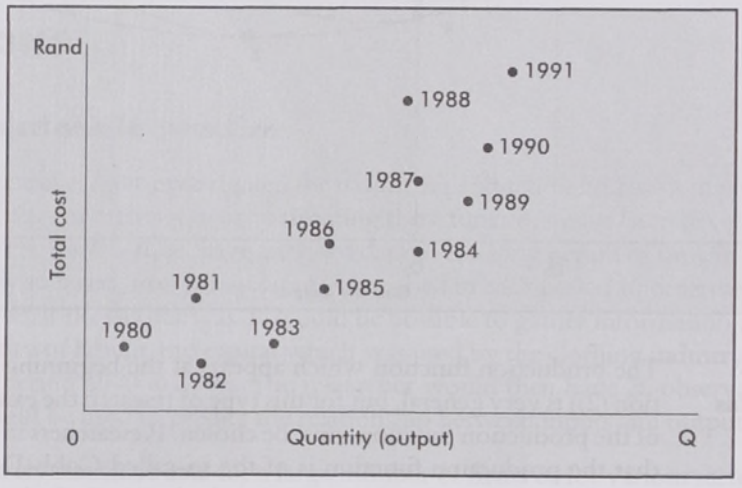
The parameters A,α and β differ from one firm to the next and also from one industry to the next.

The labour coefficient α in the Cobb-Douglas production function is the percentage increase in output which results from a 1% increase in labour, while the quantity of capital is held constant. Accordingly β is the percentage increase in output which will occur as a result of an increase of 1% in capital, when the labour input remains constant. Research of this nature, for example, would make it possible to predict for the sugar industry that if $\alpha = 0{,}59$, then an increase of 1% in the labour input would lead to a 0,59% increase in output. Accordingly, if $\beta = 0{,}33$, then an increase of 1% in the capital input will result in sugar production increasing by 0,33%. It is therefore possible to compare various industries and determine where changes in capital or labour will make the greatest changes to output.

Figure 4–17

The relationship between total costs and output of a specific firm

The levels of output of a firm are plotted against the total costs over a number of years. This gives a rough approximation of the firm's cost function.



Estimation of costs

Businesses which expand (or become smaller) are usually interested in how costs will change if output varies. Estimations in terms of future costs can be obtained by calculating a cost function which shows the relationship between costs of production and the level of output (and possibly the relationship between costs and other variables which the firm can control). As with production functions, economists have carried out many studies to calculate cost functions (or cost

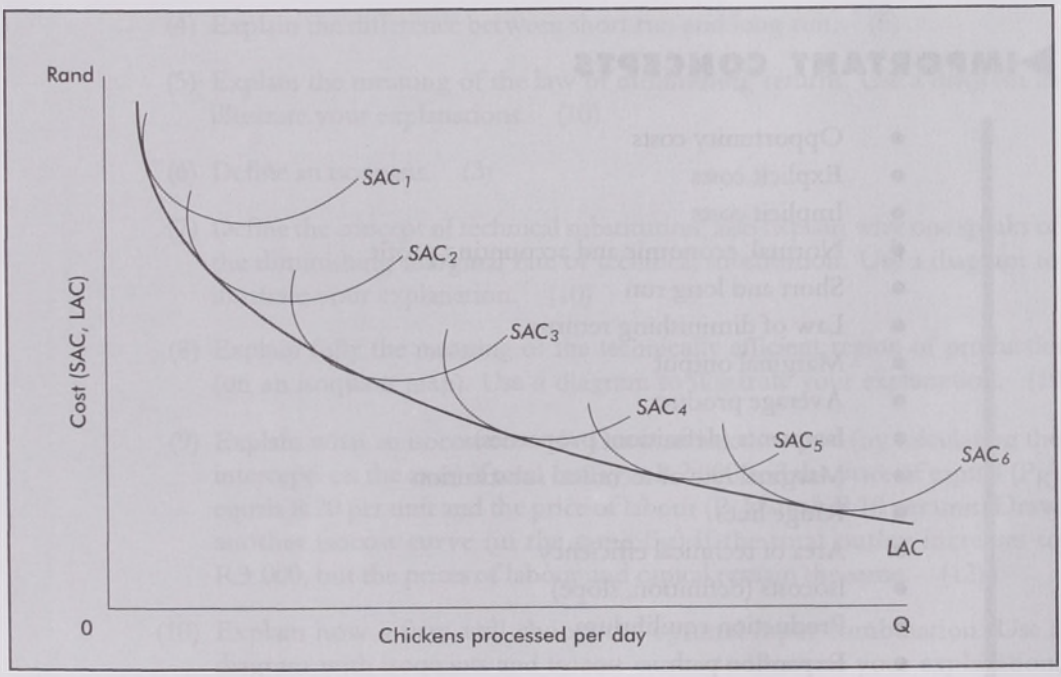
curves, as they are often called) for firms and industries. These studies are usually also based on the statistical analysis of historical data given on costs and output. (It should be clear that this type of information goes hand in hand with what was discussed in the previous section – there the relationship between inputs and output was studied, here it is the relationship between total costs and output.)

As has been mentioned, typical studies of this nature are based on historical time series data. An example of this appears in Figure 4–17, where the output levels of an imaginary firm are plotted against the total costs over a number of years. In order to predict costs relatively accurately, it is necessary to calculate the underlying relationship between costs and output. This is done with the aid of statistical methods, whereby the long-run total cost curve is estimated with the available data. If a firm wants to increase production, the accompanying costs can be predicted. This can therefore be an important aid in decision making.

Figure 4–18

The construction of the LAC curve with the aid of SAC curves

The fact that the LAC curve envelopes the SAC curves provides a method of determining whether economies of scale exist.



Construction of LAC curves

Earlier in the chapter the long-run total cost curve was used to derive the long-run average cost curve (see Fig 4-8). We also saw that if the SAC curves for different plant sizes of a firm are available, the LAC curve forms an envelope around them (see Fig 4-16). This is not only of academic interest – the phenomenon of an envelope curve allows economists to calculate the LAC curve in an alternative manner, which in practice is often the easiest way of determining whether economies of scale can be achieved.

Envelope curve

For example, a few decades ago it became obvious that the price of chicken broilers could be decreased by increasing the size of the plants where the chickens were slaughtered and processed. In one such research project the short-run average cost curves of a number of plant sizes were determined – they varied from plants that processed a few hundred chickens per day to plants that processed thousands. Figure 4-18 provides an indication of how the SAC curves are used to determine the LAC curve (or envelope curve). It is obvious that if the SAC curves are known it is not difficult to draw the LAC curve. In conclusion it can be noted that the LAC curve in Figure 4-18 corresponds with what was said previously, namely that modern theory on LAC curves claims that the curves are L-shaped (instead of being U-shaped). This indicates that economies of scale can be achieved up to very high levels of output.

IMPORTANT CONCEPTS

- Opportunity costs
- Explicit costs
- Implicit costs
- Normal, economic and accounting profit
- Short and long run
- Law of diminishing returns
- Marginal output
- Average product
- Isoquants (definition, properties)
- Marginal rate of technical substitution
- Ridge lines
- Area of technical efficiency
- Isocosts (definition, slope)
- Production equilibrium
- Expansion path
- Long-run total cost (derivation)
- Long-run average cost (definition, derivation)

- Long-run marginal cost (definition, derivation)
- Economies and diseconomies of scale
- Short-run total cost (derivation)
- Total constant cost
- Total variable cost (definition, derivation)
- Average constant cost (definition, derivation)
- Average variable cost (definition, derivation)
- Short-run average cost (definition, derivation)
- Short-run marginal cost (definition, derivation)
- Relationship between short-run and long-run cost

QUESTIONS

(1) Explain fully the concept of opportunity cost. (5)

(2) Explain what is meant by normal profit and economic profit. (6)

(3) Explain why accounting profit is greater than economic profit. (5)

(4) Explain the difference between short run and long run. (6)

(5) Explain the meaning of the law of diminishing returns. Use a diagram to illustrate your explanations. (10)

(6) Define an isoquant. (3)

(7) Define the concept of technical substitution; also explain why one speaks of the diminishing marginal rate of technical substitution. Use a diagram to illustrate your explanation. (10)

(8) Explain fully the meaning of the technically efficient region of production (on an isoquant map). Use a diagram to illustrate your explanation. (10)

(9) Explain what an isocost curve is. Draw an isocost curve (by calculating the intercepts on the axes) if total outlay is R2 000 and the price of capital (P_K) equals R20 per unit and the price of labour (P_L) equals R10 per unit. Draw another isocost curve (in the same fig) if the total outlay increases to R3 000, but the prices of labour and capital remain the same. (12)

(10) Explain how a firm will choose the optimal input combination. Use a diagram with isoquants and isocost curves to illustrate your explanation. Also explain how the equation for production equilibrium, $MP_L/P_L = MP_K/P_K$, is derived and the meaning thereof. (15)

(11) Explain how the long-run expansion path of a firm is derived with the aid of isoquants and isocost curves. Use a diagram to illustrate your explanation. (12)

(12) Explain in your own words (with the aid of a diagram) how isoquants, isocost curves and an expansion path can be used to derive the LTC curve for a firm. Draw the LMC and LAC curves that correspond with the LTC curve. (15)

(13) Use a diagram representing an LTC curve and explain how this curve can be used to derive the LMC and LAC curves (you must also draw the last-mentioned two). (10)

(14) Define the concepts of economies and diseconomies of scale. Also explain how they influence the shape of the LAC curve. (8)

(15) Draw a diagram representing the STC and TCC curves. Then explain how the figure can be used to derive the TVC curve. (You must also draw the TVC curve.) (5)

(16) Explain how the so-called 'envelope' curve can be derived if (say) three SAC curves of different size plants are available. Use a diagram to illustrate your explanation. (8)