Chapter 12: Cost Curves

12.1: Introduction

In chapter 11 we found how to minimise the cost of producing any given level of output. This enables us to find the cheapest cost of producing any given level of output. This chapter studies this cheapest cost and explores its properties. In chapter 13 we use these properties to find the profit-maximising output of the firm.

Let us denote by \( y \) some level of output of the firm. We denote by \( C(y) \) the minimum cost of producing that output. We call this function \( C(.) \) the firm’s cost function and a graph of it a cost curve.

We should note that this cost function depends upon various things, which we could include in the notation but we leave implicit. The key things that \( C(.) \) depend upon are:

1) the technology of the firm
2) the prices of the two inputs.

We should also note that there may be different cost functions – depending upon the constraints under which the firm is operating. Economists find it useful to distinguish between two scenarios (which determine the constraints under which the firm is operating), which are referred to as the long run and the short run. The long run is the familiar case in which the firm is free to vary the quantities of the two inputs. The short run is defined to be a situation in which just one of the two inputs is freely variable, while the other is fixed; we take input 1 to be freely variable in the short run while input 2 is fixed, its value \( q_2 \) equal to some fixed level \( Q_2 \). If it helps, you can imagine input 1 to be labour while input 2 is capital: in the short run the firm can not change the level of its capital – only its labour; however in the long run it can vary both.

So we can have a long-run cost function and a short-run cost function. Obviously they are different – in the short run we have this additional constraint that input 2 is fixed. This implies that the short-run cost function is different from the long-run cost function – you should be able to work out that the short-run function can never be lower than the long-run function. (We are minimising something – if we minimise something with a constraint the minimum must be at least as large as the minimum without the constraint.) We will give examples during this chapter.

I should emphasise that \( C(y) \) measures the minimum total cost of producing the output \( y \). Later in the chapter we will derive from this total cost function two other cost functions – the marginal cost function – which measures the rate at which the total costs are increasing – and the average cost function – which measures the average cost of producing a particular output.

12.2: The Long Run Total Cost Curve

We start with the long run – in which the firm can freely choose both the level of input 1 and the level of input 2. I find it useful to start with a specific example – which shows the important properties I want to discuss. Let us take the case of Cobb-Douglas technology. This has a
production function given by \( y = A q_1^a q_2^b \). In chapter 11 we found the cost-minimising input combinations for any given level of output. These are given in equation (11.3) and are:

\[
q_1 = (y/A)^{1/(a+b)}(aw_2/(bw_1))^{(b/(a+b))} \quad \text{and} \quad q_2 = (y/A)^{1/(a+b)}(bw_1/(aw_2))^{(a/(a+b))}
\]

Now the cost of using the combination \((q_1, q_2)\) is obviously \( w_1q_1 + w_2q_2 \) so the cheapest cost of producing output \( y \) in the long run is found by substituting in the optimal input demands into the production function. This yields:

\[
C(y) = (y/A)^{1/(a+b)}[w_1(aw_2/(bw_1))^{(b/(a+b))} + w_2(y/A)^{1/(a+b)}(bw_1/(aw_2))^{(a/(a+b))}]
\]

This is the long run total cost function in the Cobb-Douglas case. It can be simplified a little to give:

\[
C(y) = (a+b)(y/A)^{1/(a+b)}[w_1^{a/(a+b)}w_2^{b/(a+b)}]
\]  

(12.1)

Let us take a particular numerical example. Put \( A = 1, a = 0.3 \) and \( b = 0.5 \). This technology exhibits decreasing returns to scale. (Why? Since \( a + b = 0.8 < 1 \).) If we substitute these numbers in the expression above we get

\[
C(y) = 1.938 y^{1.25} w_1^{0.375} w_2^{0.625}
\]

We note that this is an increasing convex function of \( y \) and an increasing concave function of the two input prices. If we graph the cost function against \( y \) we get the following:

 Ensure that you understand the connection between the fact that the technology displays decreasing returns to scale and the fact that the cost function is convex: as output rises the scale has to rise proportionately faster, which makes the cost rise proportionately faster. Note from (12.1) this is a general property. In fact we should write this as a result. It is very important.

The total long run cost function is concave, linear or convex according as the technology displays increasing, constant or decreasing returns to scale.

12.3: The Short Run Total Cost Curve

Before we examine an example, we should think a little about this curve. Let me remind you of the scenario. Input 2 is fixed at the level \( Q_2 \). The firm cannot change this, It can, however, vary input 1. It still wants to do what it has been doing: for any given level of output it wants to find the cheapest way of producing that output – and thus it wants to find the cheapest cost of producing each level of
output. Do you think that with this additional constraint (it cannot vary the amount of input 2) it can do better or worse than without it?

The answer I hope is immediate: with this additional constraint it is almost always doomed to do worse than before. Therefore the short run total cost function can never be lower than the long-run total cost function – if it were then there would be a contradiction – the long run function could not be minimising the cost in the long run. However, we could be lucky in the sense that the fixed quantity of input 2 might just be the right amount to minimise cost in the long run – in such a case the long and the short run costs would coincide.

Let us now take an example. Let us continue to use the Cobb-Douglas technology we used in section 12.2 \( y = A q_1^a q_2^b \). Now additionally we have \( q_2 = Q_2 \). So we have

\[ y = A q_1^a Q_2^b \]

On the right-hand side the only thing that is variable is \( q_1 \) – notice as we vary it then the output \( y \) varies. If we want to produce a particular level of output then it is clear that there is a unique value of \( q_1 \) for which this is possible – the one that solves the above equation in terms of \( y \). Solving it we get

\[ q_1 = (y/A)^{1/a} Q_2^{-b/a} \]

To produce the output \( y \) this and only this quantity of input 1 should be used. (Note that, rather obviously, it depends upon \( Q_2 \).) It follows that the total cost of producing output \( y \) in the short run is given by \( w_1 q_1 + w_2 Q_2 \) where \( q_1 \) is given by the expression above. In this way we get the total short run cost function:

\[ C(y) = w_1 (y/A)^{1/a} Q_2^{-b/a} + w_2 Q_2 \quad (12.2) \]

We note that it is an increasing convex (if \( 0 < a < 1 \)) function of \( y \), that it starts at the value \( w_2 Q_2 \) when \( y \) is zero (this is the cost of the fixed factor), and that it is increasing in both the input prices.

Let us graph it. Obviously its position depends upon the value of \( Q_2 \) – the value at which input 2 is fixed. We give several examples, starting with \( Q_2 = 50 \). In the example that follows, as indeed in the figure above, I have taken \( w_1 = w_2 = 1 \), so the fixed cost\(^1\) is 50. Hence the graph starts at 50. Thereafter it is an increasing convex function of \( y \). It is the upper graph in this figure.

\[ \text{12.2. one short run (total) cost curve} \]

You will see that this is in agreement with our prior speculations. You will also see that there is a point at which the two curves coincide (indeed they are tangential at this point). This is an output a

\(^1\) The cost of the fixed input – namely \( w_2 Q_2 \).
little below 20. What is the significance of this output? It is the output at which the employment of 50 units of input 2 would be optimal in the long run. So at that point the constraint is not really binding and the firm does as well in the short run as it does in the long.

If we take a different value of \(Q_2\) we get a different short run total cost curve. For example, with \(Q_2 = 130\) we get:

![Graph showing the short run total cost curve for different values of \(Q_2\).](image)

Note that this starts at 130 and is tangential to the long run curve at a higher level of output. Note that both short run curves have the same property – they start positive (the fixed costs) are increasing and convex and are everywhere above the long run curve except at the one point where they are tangential. We also note an important property of all the short run curves put together. Here we just plot a selection but it should be clear what would happen if we plotted more.

![Graph showing the envelope property of short run curves.](image)

As mathematicians would say: the long run curve is the *envelope* of all the short run curves.

**12.4: Marginal and Average Costs**

So far we have been talking solely about *total* costs – whether in the short run or the long. Now we want to introduce two new cost curves (for both the long and the short run) which we can derive from the total cost curve and which will prove useful to us in the future. These are the *marginal cost curve* and the *average cost curve*.

The *marginal cost curve* just tells us the rate at which total costs are increasing. Mathematically, marginal costs are the derivative of the total cost function – that is \(dC(y)/dy\). Graphically, marginal costs are the slope of the total cost function.
The **average cost curve** measures the cost per unit produced. Mathematically, average costs are the ratio of total cost to output – $C(y)/y$. Graphically – and this is a useful interpretation – average costs are the slope of the line joining the origin to the total cost curve.

We can derive marginal and average costs both in the long run and in the short run. Let us start with the former. If we look at the graph of the total cost in the long run – figure 12.1 above – we see that the slope of the curve is everywhere increasing (it is a convex function) and moreover that the slope of the line from the origin to the curve is also increasing everywhere. It follows that both marginal cost and average cost are increasing everywhere. We also note that the slope of the curve is always greater than the slope of the line from the origin to the curve – so that marginal cost is always larger than average cost. If we graph the marginal and average cost in a figure we get the following, in which the upper curve is the marginal cost curve and the lower curve the average cost curve.

Let us find the marginal and average curves for one of the short runs. Take the example pictured in figure 12.2 – here $Q_2 = 50$. Recall that the short run curve is the upper one. If we look at it we first note that the slope of it is everywhere increasing. It therefore follows that the short run marginal cost curve is everywhere increasing. Not so the average cost. In fact when $y$ is zero the average short run cost is infinite – it then decreases until an important point, which we now define. This is the point on the short run total cost curve where the line from the origin is tangential to the curve. Clearly at this point the marginal cost is equal to the average cost. Moreover at this point the average cost is at its lowest point – up to then the average cost decreases, after then it increases. This explains the following:
The ‘important point’ is an output just below 18. At this output, in figure 12.2, the line from the origin to the curve is tangential to that curve at that point. Hence in figure 12.7, the marginal cost curve intersects the average cost curve, and, moreover, the average cost curve reaches its minimum at that point. Obviously these properties are true for any short run – a proof is supplied in the Mathematical Appendix - and in particular the second of the short runs that we considered.

Note that with a higher level of the fixed input, the minimum value of the average short run cost curve occurs at a higher output level and also takes a higher minimum value.

Finally we can illustrate a second envelope property. We have already established that the long run total cost curve is the envelope of the short run cost curves. It follows from this that the long run average cost curve must be the envelope of the short run average cost curves, as the following figure shows:
Throughout the above analysis we worked with a specific example in which the technology displayed decreasing returns to scale, but the results we have obtained apply also to increasing and constant returns to scale. The only thing that really differs is the shape of the long run function — all the other properties concerning the relationship between the long and the short run curves remain valid. We know that the long-run curve is concave, linear or convex according as the technology displays increasing, constant or decreasing returns to scale. Let us take an increasing returns example: Cobb-Douglas with parameters $a = 0.45$ and $b = 0.75$. The long run cost curve is:

In the short run we almost always have decreasing returns to the variable factor — so the short run total cost functions are convex. We get the first envelope property:

---

2 In the case of the Cobb-Douglas if we put $0 < a < 1$ and $0 < b < 1$ this guarantees decreasing returns to each factor. In practice this is overwhelmingly the usual case.
Note that the long run curve is concave but each of the short run curves are convex. This reflects increasing returns to scale in the long run but decreasing returns to the variable factor in the short run. Note that because of the shape of the long run function, both the marginal and average long run cost curves are decreasing – with the marginal always less than the average:

However each of the short run marginal and average curves have the same properties as before (the marginal curve is the one that is everywhere increasing):

and we have the usual second envelope property:
Obviously everything works also with a constant returns to scale technology – the only thing that differs is that the long run total cost function is *linear* and hence that the long run average and marginal cost curves are horizontal and coincide.

### 12.6: From Marginal to Total

We have already explained how to derive a marginal cost curve from a total cost curve – we simply find the *slope* of the total cost curve. This tells us the rate at which total costs are increasing – and hence gives us marginal costs.

To a mathematician, what we are doing is finding the *derivative* of the total cost function – this tells us the marginal cost function.

This final section asks how we might go the other way – how we might derive the total cost function from knowledge of the marginal cost function.

To a mathematician, the reverse of finding the derivative is finding the integral (see the Mathematical Appendix to Chapter 1). So, if the marginal cost function is the derivative of the total, then the total is the integral of the marginal.

To a non-mathematician, an integral is an *area*. So what the above sentence is saying is the following: if the marginal cost function is the slope of the total cost function, then the total cost function is the *area underneath* the marginal cost function. (To get the total cost we just add up all the marginal costs.) More specifically the total cost at a particular output level $y$ is given by the area underneath the marginal cost function from $\theta$ to $y$.

We provide an example from which you should be able to verify this result. Here we have in the one figure the long run total and marginal cost curves for the first of the examples in this chapter. Rather obviously the upper curve is total cost and the lower curve marginal cost.
You should work both ways. First select an output level – say 6. Draw a vertical line at this value. Then calculate the slope of the upper line at this value. Confirm that this is equal to the level of the marginal cost curve. Now do the opposite: work out the area between 0 and 6 underneath the marginal cost curve. Confirm that this is equal to the level of the total cost curve.

The same works in the short run – with one slight difference. Consider the figure below which shows total and marginal short run costs for the first of the short runs we considered in this chapter.

The top graph is the short run total cost curve; the bottom curve is the short run marginal cost curve. To go from the top curve to the bottom curve we simply find the slope of the top curve. You should check this. To go the other way – we find the area under the bottom curve. However, in this instance we notice that there is one thing missing – the fixed cost (of 50). So the area under the bottom curve is the total short run cost excluding the fixed cost. You should check this.
12.7: Summary

This chapter has provided some useful material for future chapters. You should remember and make sure that you understand the important results – which are summarised here.

We started with making a distinction between the long run and the short. In the long run the firm can freely choose the level of both inputs; in the short run, one of the two inputs is fixed. We started with the long run and used the results of chapter 11 to find the cheapest way of producing any given level of output. We defined the total long run cost curve and found an important property of the curve.

The long run curve shows the cheapest cost of producing each output. It is concave, linear or convex according as the technology shows increasing, constant or decreasing returns to scale.

We then examined the short run cost curve, and found the relationship between the short and long run curves. In particular we found that the short run curve is everywhere above the long run curve except at one point where it is tangential – this is the output level for which the fixed input is at its optimal long term value. We then have:

The long run total cost curve is the (lower) envelope of the short run total cost curves.

We then defined marginal and average costs.

The marginal cost curve shows the rate at which total costs are increasing. It must be positive everywhere. In the long run it is decreasing, constant or increasing according as the technology shows increasing, constant or decreasing returns to scale. In the short run it is usually increasing everywhere as there are usually decreasing returns to the fixed factor.

The average cost curve shows the cost per unit of output.

The average cost curve and the marginal cost curve intersect at the minimum point of the average cost curve.

From the envelope result above it follows that:

The long run average cost curve is the (lower) envelope of the short run average cost curves.

We define the fixed costs as those that are fixed in the short run and hence are the cost of producing zero output in the short run. Finally we show how to go from the marginal cost curve to the total cost curve.

The area under the long run marginal cost curve is the long run total cost.

The area under the short run marginal cost curve is the variable (total) cost.
12.8: How do we infer the technology of a firm?

This is a simple game that could be set as a tutorial exercise and which you could play with your fellow students of microeconomics. It is very close in spirit to the exercise at the end of Chapter 7, though there is an added twist. Although simple, it teaches us a lot about how we can infer technologies from observations, and the most efficient way of so doing.

This is how to organise the game. First split the tutorial group up into two teams and give to each team the following instructions.

The objective is to discover the technology of a firm given observations on (minimum) costs of producing given outputs at given factor prices. Each team should imagine themselves as a firm with a given technology. The objective of the rival team is to discover that technology. As in the exercise of chapter 7 we make the task easier by restricting the set of possible technologies.

Underlying the technology are two factors (inputs): input 1 and input 2. Input 2 is the *numéraire* input - its price is 1. The price of input 1 is $p$. This price is constant, independent of the amount of the factors used.

Each team should decide what its technology is. Technology is restricted to *one* of the following:

**Type 1: Perfect Substitutes** with substitution ratio $a$ (that is, 1 unit of input 1 can always be substituted by $a$ units of input 2), and with returns to scale indicated by the parameter $s$ - that is, $q_1$ units of input 1 and $q_2$ units of input 2 produces $(q_1 + q_2)^s$ units of output. So the production function is as given by equation (10.6), namely $y = f(q_1, q_2) = (q_1 + q_2)^s$.

**Type 2: Perfect Complements** with complementary ratio $a$ (that is, if $q_1$ units of input 1 are employed then output increases with units of input 2 until $aq_1$ units of input 2 are employed, after which no further increases in output are produced) and with returns to scale indicated by the parameter $s$ - that is, $q_1$ units of input 1 combined with $q_2$ units of input 2 produce $[\min(q_1, q_2/a)]^s$ units of output. So the production function is that given by equation (10.11), namely $y = f(q_1, q_2) = [\min(q_1, q_2/a)]^s$.

**Type 3: Cobb-Douglas** with relative weights $a$ and $(1-a)$ on the two inputs and with returns to scale indicated by the parameter $s$. So that the production function is given by a slight variant on equation (10.12), namely $y = f(q_1, q_2) = [q_1^a q_2^{(1-a)}]^s$.

Each team should choose *one* of these 3 Types and values for the parameters $a$ and $s$. *You should inform your tutor of these choices*: he or she will act as a referee in what follows.

Both teams know their own technology - their type and their values of $a$ and $s$. The purpose of the exercise is for each team to discover the technology of the other team - that is, its Type and its values for $a$ and $s$ - by asking questions of the form:

“What is your minimum cost of producing $y$ units of output when the price of input 1 is $p$?”

Only questions of this type are allowed, but $y$ and $p$ can be any positive numbers.
The way to implement the game is as follows. First both of the teams tell the tutor their Type and the value of their parameters $a$ and $s$. Ideally the teams will have decided these things before the tutorial begins. Then put on the whiteboard two tables side by side as follows:

<table>
<thead>
<tr>
<th>Team 1’s costs</th>
<th>Team 2’s costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>$p$</td>
</tr>
<tr>
<td>$Y$</td>
<td>$y$</td>
</tr>
<tr>
<td>$C(y)$</td>
<td>$C(y)$</td>
</tr>
</tbody>
</table>

The game is played simultaneously by the two teams. First both teams write up a question (a value of $p$ and a value of $y$) in the appropriate table on the whiteboard. (Team 1 in the table headed “Team 2’s cost” and Team 2 in the table headed “Team 1’s costs”). Usually it takes some time for the teams to think up this initial question, but on reflection it becomes obvious that it does not really matter too much what the teams choose for $p$ and $y$. Then the teams have to fill in their answers – the cost as given by their technologies at the $p$ and $y$ values selected by the other team. This stage takes a bit of time, unless the teams are really well prepared. I, as the referee, check that the answers are consistent with the technologies of the two teams. This is important as students frequently make mistakes.

At this point we have one complete row in each of the above tables. The teams are then invited to try and infer the technologies of the other team: their Type and the values of the parameters $a$ and $s$. I do not want to tell you too much at this point (otherwise it will spoil the point of the game) but only a certain amount can usually be inferred. Unless the cost is close to zero for a $p$ value close to zero, one can not eliminate any Type of technologies. However, one can usually conclude something like the following: the technology could be perfect substitutes with particular values of the parameters $a$ and $s$ (determined by the question and the answer); the technology could be perfect complements with particular sets of values of the parameters $a$ and $s$ (determined by the question and the answer); the preferences could be Cobb-Douglas with particular sets of values of the parameters $a$ and $s$ (determined by the question and the answer).

Since nothing definite has been inferred by this stage, we repeat the procedure. The teams are invited to write up a second question (a value of $p$ and a value of $y$) in the appropriate table on the whiteboard. The choice of $p$ and $y$ in this second question has to be done carefully. In this exercise it is quite clever to use the same value of $p$ as in their first question, and change the value of $y$. As is discussed in the hints below you can use the information gained to determine the returns to scale parameter $s$. Once this is known it is quite easy then to discover the Type of technology, and, though possibly less easily, the value of the parameter $a$.

This whole procedure is repeated until one Team infers the Type of the other and the values of the parameters $a$ and $s$.

Some hints are given below, but I would recommend you do not read these hints until you have played the game. You will learn more if you work through the game yourselves.

You could, of course, make the game more interesting and more difficult by allowing further possible types of technologies. One obvious extension is to include Constant Elasticity of Substitution technology. This makes it more difficult – but more realistic. In practice we do not have a set of technologies from which to choose, so the set of possible technologies could be very large. There are other types of technologies that economists use, and which are better
approximations to actual technologies used by actual firms, but discussing these would take us beyond the scope of this book.

12.9: Mathematical Appendix

We start with a proof of the result that the marginal cost curve intersects the average cost curve at the minimum point of the average cost curve. To do this we note that the average cost curve is simply:

\[ AC(y) = \frac{C(y)}{y} \]

To find the value of \( y \) where this is minimised we find the point where \( dAC(y)/dy = 0 \), that is the bottom of the average cost curve. If we evaluate \( dAC(y)/dy \) we get, using the quotient rule for derivatives which we noted in Chapter 1:

\[ dAC(y)/dy = \frac{dC(y)/dy}{y} - C(y)/y^2 \]

and hence \( dAC(y)/dy = 0 \) where \( \frac{dC(y)/dy}{y} - C(y)/y^2 = 0 \), that is, where \( dC(y)/dy = C(y)/y \), that is, where marginal cost is equal to the average cost. Hence our result.

12.10: Hints Appendix

Hints and comments – which you are recommended to read AFTER you have played the game.

This is quite a difficult, but very relevant, exercise. It is relevant because the whole point of this exercise is to show how we can infer technology from observations on costs. Once we have this information we can use it to predict the future behaviour of the firm. As for the difficulty of the exercise, it all depends on how you tackle it. Use the knowledge that is contained in your head and the text concerning the relation between the minimum production cost and the price of the input \( p \) and the output of the firm. Note that in general the form of these relationships depends upon the technology, though there is an important result that you can exploit. Consider changing the value of \( y \) without changing the value of \( p \). The isocost curves keep the same slope. Given that the isoquants in all the technologies considered here have the property that they are radially parallel\(^3\) (which means that, along any ray from the origin, the slope of every isoquant is the same) it follows that the quantities of both inputs increase in the same proportion. Suppose we multiply the output required by \( c \) and we find that the factor inputs are both multiplied by \( d \). Then we have \( y = f(q_1,q_2) \) and \( cy = f(dq_1,dq_2) \). But, for all three technologies that we are considering, (see equations (10.6), (10.11) and (10.12) above), it is true that \( f(dq_1,dq_2) = d^s f(q_1,q_2) \), where \( s \) is the returns to scale parameter. Putting these equations together we have that \( c = d^s \). Now we know \( c \) because we have specified that, and we know \( d \) because that is simply by how much the cost has been multiplied. So we can work out the returns to scale parameter \( s \).

I do not want to say too much more, but you should use your intuition. We have already used the relationship between the cost and \( y \) to infer the value of \( s \). To find the technology and the value of the parameter \( a \), we need to use our knowledge about the relationship between the cost and \( p \). Consider for example Type 1 – perfect substitutes. What do you we know? That the firm uses input

\[^3\] Formally this property is known as homotheticity of the isoquants.
1, and only input 1, when it is cheap enough \((p < a)\) and uses input 2, and only input 2, when the input 1 is too expensive \((p > a)\). So if we plot the cost against \(p\), it rises with \(p\) until \(p\) gets to \(a\), and then it remains horizontal (as a function of \(p\)). If this is what we find when we graph the cost against \(p\), we can infer that the technology is Type 1 and the value of \(a\) is the price at which the graph stops rising. Similar arguments – which we leave to you – apply for the other two types.