# FACULTY OF BUSINESS AND ECONOMICS

# DEPARTMENT OF MANAGEMENT

## *MANAGERIAL ECONOMICS*

*(*MGMT 3017*)*

**ADDIS ABABA, ETHIOPIA**
**NOVEMBER, 2023**

# Contents

# CHAPTER ONE
# 1. SCOPE AND NATURE OF MANAGERIAL ECONOMICS

***Objective***

- *To introduce and define managerial economics*
- *To explain the nature and scope of managerial economics*
- *To outline the types of issue which are addressed by managerial economics*

## 1.1. What is Managerial Economics?

Managerial economics is the application of economic theory and quantitative methods (mathematics and statistics) to the managerial decision-making process. Simply stated, managerial economics is applied microeconomics with special emphasis on those topics of greatest interest and importance to managers. Managerial economics is the synthesis of microeconomic theory and quantitative methods to find optimal solutions to managerial decision-making problems.

Managerial economics applies economic theory and methods to business and administrative decision making. Managerial economics prescribes rules for improving managerial decisions. Managerial economics also helps managers to recognize how economic forces affect organizations and describes the economic consequences of managerial behaviour.

Managerial economics refers to the application of economic theory and the tools of analysis of decision science to examine how an organization can achieve its aims or objectives most efficiently. It links traditional economics with the decision sciences to develop vital tools for managerial decision making. Managerial economics identifies ways to efficiently achieve goals. For example, suppose a small business seeks rapid growth to reach a size that permits efficient use of national media advertising. Managerial economics can be used to identify pricing and production strategies to help meet this short-run objective quickly and effectively. Many different definitions have been given but most of them involve the application of economic theory and methods to business decision-making. As such it can be seen as a means to an end by managers, in terms of finding the most efficient way of allocating their scarce resources and reaching their objectives.

As an approach to decision-making, managerial economics is related to economic theory, decision sciences and business functions. These relationships can be discussed.

- **Relationship with economic theory**

The main branch of economic theory with which managerial economics is related is microeconomics, which deals essentially with how markets work and interactions between the various components of the economy. In particular, the following aspects of microeconomic theory are relevant:

- theory of the firm
- theory of consumer behaviour (demand)
- production and cost theory (supply)
- price theory
- market structure and competition theory

**Relationship with decision sciences**

The decision sciences provide the tools and techniques of analysis used in managerial economics. The most important aspects are as follows:

- numerical and algebraic analysis
- optimization
- statistical estimation and forecasting
- analysis of risk and uncertainty
- discounting and time-value-of-money techniques

These tools and techniques are introduced in the appropriate context, so that they can be immediately applied in order to understand their relevance, rather than being discussed in block or in isolation.

**Relationship with business functions**

All firms consist of organizations that are divided structurally into different departments or units, even if this is not necessarily performed on a formal basis. Typically, the units involved are:

- production and operations
- marketing
- finance and accounting
- human resources

All of these functional areas can apply the theories and methods mentioned earlier, in the context of the particular situation and tasks that they have to perform. Thus a production department may want to plan and schedule the level of output for the next quarter, the marketing department may want to know what price to charge and how much to spend on advertising, the finance department may want to determine whether to build a new factory to expand capacity, and the human resources department may want to know how many people to hire in the coming period and what it should be offering to pay them. It might be

noted that all the above decisions involve some kind of quantitative analysis; not all managerial decisions involve this kind of analysis. There are some areas of decision-making where the tools and techniques of managerial economics are not applicable. The meaning of this relationship can be seen in the figure below.

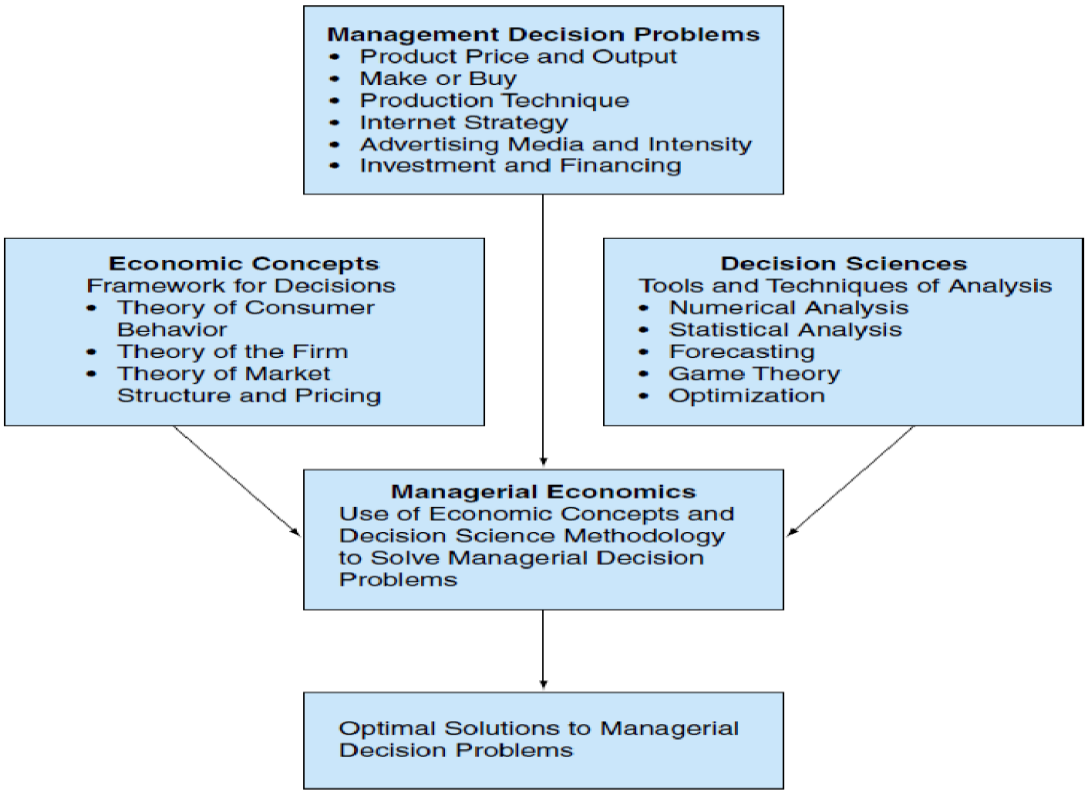


*Figure 1.1. Nature of Managerial Economics*

### 1.2. Scope of Managerial Economics

Two major managerial functions served by subject matter under managerial economics are;

**i) Decision making**

a. *Decision related to demand* – a manager has to take decisions regarding the quality and quantity of his product. For taking such decisions, a number of economic tools can be applied. Such as demand, elasticity of demand, demand forecasting etc.

b. *Decisions related to cost and production* – a manager has to analyse the cost and the types of cost he has to incur in the production process. He has to know the different laws governing production.

c. *Decisions related to price and market* – market analysis is part of managerial economics, as a manager should have the knowledge of various market structures and how price output is determined under the different types of market structure.

d. *Decision related to profit management* – some policies and theories related to profit management should be included.

e. *Macroeconomic factor* – a firm never works in isolation, so understanding of macro level factors by any business man is important. E.g. business cycle and inflation

**ii) Forward planning –** the other managerial function served by managerial economics is planning in which different economic theories like demand forecasting help the manager to appropriately plan for the firm.

***Self Assessment Exercises***

1. What is the relationship between the field of managerial economics and

(a ) Economic theories like microeconomics and macroeconomics?

( b) Decision sciences?

(c) The fields of accounting, finance, marketing, human resource, and production?

2.Define Managerial Economics using your own terms.

3.Explain the scope of Managerial Economics.

4.Explain how managerial economics is similar to and different from microeconomics.

5.Why is the subject of managerial economics relevant to the problem of decision making?

# CHAPTER TWO
# 2. ECONOMIC OPTIMIZATION

***Objective***

- *Explain the meaning and types of optimization problems.*
- *Examine the relationship between marginal analysis and differential calculus.*
- *Find optimal solutions using calculus to many kinds of maximization and minimization problems in managerial economics.*
- *Solve the constrained optimization problems using Lagrangian multiplier techniques.*

## 2.1. Introduction

Effective managerial decision making is the process of arriving at the best solution to a problem. If only one solution is possible, then no decision problem exists. When alternative courses of action are available, the best decision is the one that produces a result most consistent with managerial objectives. The process of arriving at the best managerial decision is the goal of economic optimization and the focus of managerial economics.

Normative economic decision analysis involves determining the action that best achieves a desired goal or objective. This means finding the action that optimizes (i.e., maximizes or minimizes) the value of an objective function. For example, in a price-output decision-making problem, we may be interested in determining the output level that maximizes profits. In a production problem, the goal may be to find the combination of inputs (resources) that minimizes the cost of producing a desired level of output. In a capital budgeting problem, the objective may be to select those projects that maximize the net present value of the investments chosen. There are many techniques for solving optimization problems such as these. Optimization techniques are a powerful set of tools that are important in efficiently managing a firm ‘s resources and thereby maximizing shareholder wealth.

## 2.2. Types of Optimization Techniques

The basic form of the problem in decision making is to identify the alternative means of achieving a given objective and then to select the alternative that accomplishes the objective in the most efficient manner, subject to constraints on the means. In programming terminology, the problem is optimizing the value of some objective function, subject to any

resource and/or other constraints such as legal, input, environmental, and behavioural restrictions. Mathematically, we can represent the problem as

$$\text{Optimize } y = f(x_1, x_2, \ldots, x_n) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \textit{Equation 2.1}$$

$$\text{subject to } g_j(x_1, x_2, \ldots, x_n) \begin{vmatrix} \leq \\ = \\ \geq \end{vmatrix} b_j \quad j = 1, 2, \ldots, m \qquad \ldots\ldots\ldots\ldots\ldots\ldots \textit{Equation 2.2}$$

where Equation 2.1 is the objective function and Equation 2.2 constitutes the set of constraints imposed on the solution. The $x_i$ variables, $x_1$, $x_2$, . . ., $x_n$, represent the set of decision variables, and $y$ $f(x_1, x_2, \ldots, x_n)$ is the objective function expressed in terms of these decision variables.

Depending on the nature of the problem, the term *optimize* means either *maximize* or *minimize* the value of the objective function. As indicated in Equation 2.2, each constraint can take the form of an equality (=) or an inequality (≤ or ≥) relationship.  ***Complicating Factors in Optimization***

The following are several factors that can make optimization problems fairly complex and difficult to solve.

- *The existence of multiple decision variables in a problem*

Relatively simple procedures exist for determining the profit-maximizing output level for the single-product firm. However, the typical medium- or large-size firm often produces a large number of different products, and as a result, the profit maximization problem for such a firm requires a series of output decisions-one for each product.

- *The complex nature of the relationships between the decision variables and the associated outcome*

For example, in public policy decisions, on government spending for such items as education, it is extremely difficult to determine the relationship between a given expenditure and the benefits of increased income, employment, and productivity it provides. No simple relationship exists among the variables. Many of the optimization techniques discussed here are only applicable to situations in which a relatively simple function or relationship can be postulated between the decision variables and the outcome variable.

- *The possible existence of one or more complex constraints on the decision variables*

For example, virtually every organization has constraints imposed on its decision variables by the limited resources-such as capital, personnel, and facilities- over which it has control.

These constraints must be incorporated into the decision problem. Otherwise, the optimization techniques that are applied to the problem may yield a solution that is unacceptable from a practical standpoint.

- *The presence of uncertainty or risk*

The presence of uncertainty or risk illustrates the difficulties that may be encountered and may render a problem unsolvable by formal optimization procedures.

The mathematical techniques used to solve an optimization problem represented by Equations 2.1 and 2.2 depend on the form of the criterion and constraint functions. The simplest situation to be considered is the *unconstrained* optimization problem. In such a problem no constraints are imposed on the decision variables, and *differential calculus* can be used to analyze them. Another relatively simple form of the general optimization problem is the case in which all the constraints of the problem can be expressed as *equality* (=) relationships. The technique of *Lagrangian multipliers* can be used to find the optimal solution to many of these problems.

Often, however, the constraints in an economic decision-making problem take the form of *inequality* relationships ($\leq$ or $\geq$) rather than equalities. For example, limitations on the resources-such as personnel and capital-of an organization place an *upper bound* or budget ceiling on the quantity of these resources that can be employed in maximizing (or minimizing) the objective function. With this type of constraint, all of a given resource need not be used in an optimal solution to the problem. An example of a *lower bound* would be a loan agreement that requires a firm to maintain a *current ratio* (that is, ratio of current assets to current liabilities) of at least 2.00. Any combination of current assets and current liabilities having a ratio greater than or equal to 2.00 would meet the provisions of the loan agreement. Such optimization procedures as the Lagrangian multiplier method are not suited to solving problems of this type efficiently; however, modern mathematical programming techniques have been developed that can efficiently solve several classes of problems with these inequality restrictions.

*Linear-programming* problems constitute the most important class for which efficient solution techniques have been developed. In a linear-programming problem, both the objective and the constraint relationships are expressed as linear functions of decision variables. Other classes of problems include *integer programming* problems, in which some (or all) of the decision variables are required to take on integer values, and *quadratic-programming* problems, in which the objective relationship is a quadratic function of the

decision variables. Generalized computing algorithms exist for solving optimization problems that meet these requirements. The various types of optimization problems and techniques can, thus, be represented as follows:

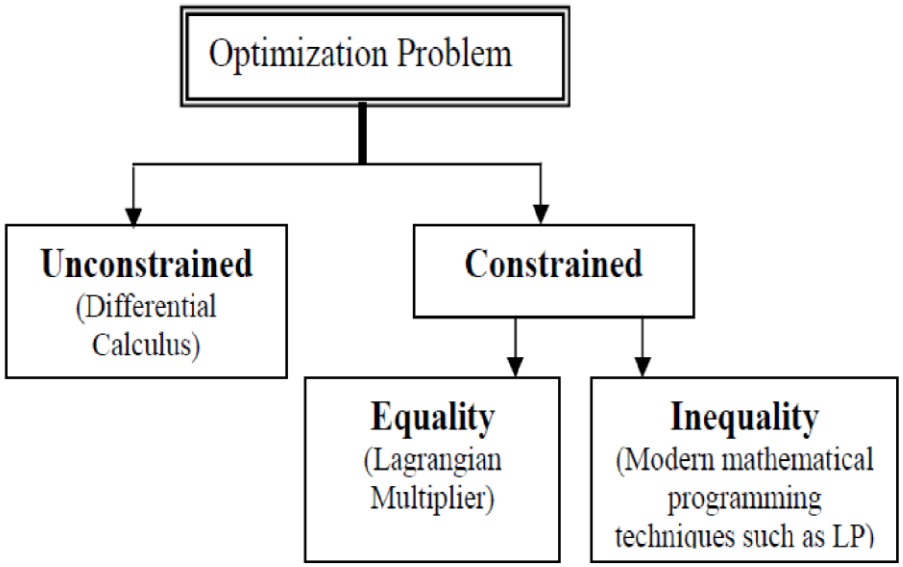


*Figure 2.1. Optimization techniques*

### 2.2.1. Differential Calculus and Bi-variate Optimization

Recall that in chapter 1, marginal analysis was introduced as one of the fundamental concepts of economic decision making. In the marginal analysis framework, resource-allocation decisions are made by comparing the marginal benefits of a change in the level of an activity with the marginal costs of the change. A change should be made as long as the marginal benefits exceed the marginal costs. By following this basic rule, resources can be allocated efficiently and profits or shareholder wealth can be maximized.

In the profit-maximization, the application of the marginal analysis principles required that the relationship between the objective (profit) and the decision variable (output level) be expressed in either tabular or graphic form. This framework, however, can become cumbersome when dealing with several decision variables or with complex relationships between the decision variables and the objective. When the relationship between the decision variables and criterion can be expressed in *algebraic* form, the more powerful concepts of differential calculus can be used to find optimal solutions to these problems

- ***Relationship between Marginal Analysis and Differential Calculus***

Initially, let us assume that the objective we are seeking to optimize, *Y*, can be expressed algebraically as a function of *one* decision variable, *X*,

$Y = f(X)$……. *Equation 2.3*

Recall that marginal profit is defined as the change in profit resulting from a one-unit change in output. In general, the marginal value of any variable *Y*, which is a function of another variable *X*, is defined as the change in the value of *Y* resulting from a one-unit change in *X*. The marginal value of *Y*, *My*, can be calculated from the change in *Y*, *ΔY, which* occurs as the result of a given change in *X*, *ΔX:*

$My = \Delta Y / \Delta X$ .....*Equation 2.4*

When calculated with this expression, different estimates for the marginal value of *Y* may be obtained, depending on the size of the change in *X* that we use in the computation. The true marginal value of a function (e.g., an economic relationship) is obtained from Equation 2.4 when *X* is made as small as possible.

If *X* can be thought of as a continuous (rather than a discrete) variable that can take on fractional values, then in calculating *M y* by Equation 2.4, we can let *X* approach zero. In concept, this is the approach taken in differential calculus. The derivative or, more precisely, *first derivative, dY/dX,* of a function is defined as the *limit* of the ratio *ΔY/ ΔX* as *ΔX* approaches zero; that is,

$$\frac{dY}{dX} = \text{limit}_{\Delta X \to 0} \frac{\Delta Y}{\Delta X}$$ .....*Equation 2.5*

Graphically, the first derivative of a function represents the *slope* of the curve at a given point on the curve. The definition of a derivative as the limit of the change in *Y* (that is, *ΔY*) as *ΔX* approaches zero is illustrated in Figure 2.3(a). Suppose we are interested in the derivative of the $Y = f(X)$ function at the point $X_0$. The derivative *dY/dX* measures the slope of the tangent line *ECD.* An estimate of this slope, albeit a poor estimate, can be obtained by calculating the marginal value of *Y* over the interval $X_0$ to *X2*. Using Equation 2.4, a value of

$$M_y' = \frac{\Delta Y}{\Delta X} = \frac{Y_2 - Y_0}{X_2 - X_0}$$

is obtained for the slope of the *CA* line. Now let us calculate the marginal value of *Y* using a smaller interval, for example, *X*0 to *X*1. The slope of the *CB* line, which is equal to

$$M_y'' = \frac{\Delta Y}{\Delta X} = \frac{Y_1 - Y_0}{X_1 - X_0}$$

gives a much better estimate of the true marginal value as represented by the slope of the *ECD* tangent line. Thus we see that the smaller the *ΔX* value, the better the estimate of the

slope of the curve. Letting *ΔX approach* zero allows us to find the slope of the $Y = f(X)$ curve at point *C*. As shown in Figure 2.3(b), the slope of the *ECD* tangent line (and the $Y = f(X)$ function at point *C*) is measured by the change in *Y*, or rise, *ΔY,* divided by the change in *X*, or run, *ΔX.*

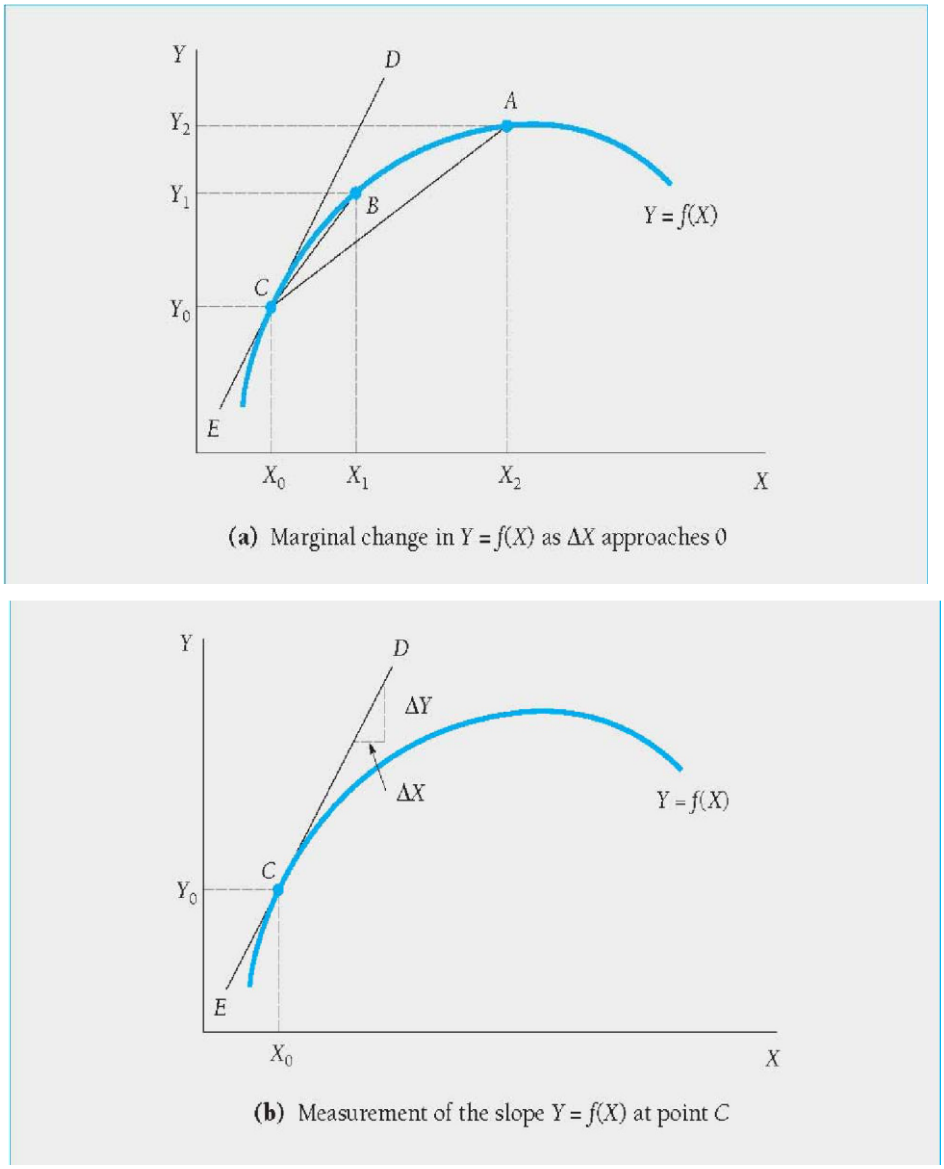


*Figure 2.2. Marginal change and slope function*

- **Process of Differentiation**

The process of differentiation-that is, finding the derivative of a function- involves determining the limiting value of the ratio *ΔY/ ΔX* as *ΔX* approaches zero. Before offering some general rules for finding the derivative of a function, we illustrate with an example the algebraic process used to obtain the derivative without the aid of these general rules.

**Example:** *Process of differentiation: profit maximization at Mekelle Power Co.*

Suppose the profit, π, of Mekelle Power Co. can be represented as a function of the output level $Q$ using the expression

$\pi = -40 + 140Q - 10Q^2$ ⋯……. *Equation 2.6*

We wish to determine $d\pi/dQ$ by first finding the marginal-profit expression $\Delta\pi/\Delta Q$ and then taking the limit of this expression as $\Delta Q$ approaches zero.

Let us begin by expressing the new level of profit ($\pi = \Delta\pi$) that will result from an increase in output to ($Q + \Delta Q$).

From Equation 2.6, we know that

$(\pi + \Delta\pi) = -40 + 140(Q + \Delta Q) - 10(Q + \Delta Q)^2$……...*Equation 2.7*

Expanding this expression and then doing some algebraic simplifying, we obtain

$(\pi + \Delta\pi) = -40 + 140Q + 140\Delta Q - 10\,[Q^2 + 2Q\,\Delta Q + (\Delta Q)^2]$

$= -40 + 140Q + 140\Delta Q - 10Q^2 - 20Q\,\Delta Q - 10(\Delta Q)^2$……. *Equation 2.8*

Subtracting Equation 2.6 from Equation 2.8 yields

$\Delta\pi = 140\,\Delta Q - 20Q\,\Delta Q - 10(\Delta Q)^2$……...……. *Equation 2.9*

Forming the marginal-profit ratio $\Delta\pi / \Delta Q$, and doing some cancelling, we get

$$\frac{\Delta\pi}{\Delta Q} = \frac{140\Delta Q - 20Q\Delta Q - 10(\Delta Q)^2}{\Delta Q}$$
$$= 140 - 20Q - 10\Delta Q$$

……*Equation 2.10*

Taking the limit of Equation 2.10 as $\Delta Q$ approaches zero yields the expression for the derivative of Mekelle Power Co. ‘s profit function (Equation 2.6)

$$\frac{d\pi}{dQ} = \lim_{\Delta Q \to 0} [140 - 20Q - 10\Delta Q]$$
$$= 140 - 20Q$$

……*Equation 2.11*

If we are interested in the derivative of the profit function at a particular value of $Q$, Equation

2.11 can be evaluated for this value. For example, suppose we want to know the marginal profit, or slope of the profit function, at $Q = 3$ units. Substituting $Q = 3$ in Equation 2.11 yields

$$\text{Marginal profit} = \frac{d\pi}{dQ} = 140 - 20(3) = \$80 \text{ per unit}$$

- **Rules of Differentiation**

Fortunately, we do not need to go through a lengthy process every time we want the derivative of a function. A series of general rules exists for differentiating various types of functions as shown in figure 2.4.

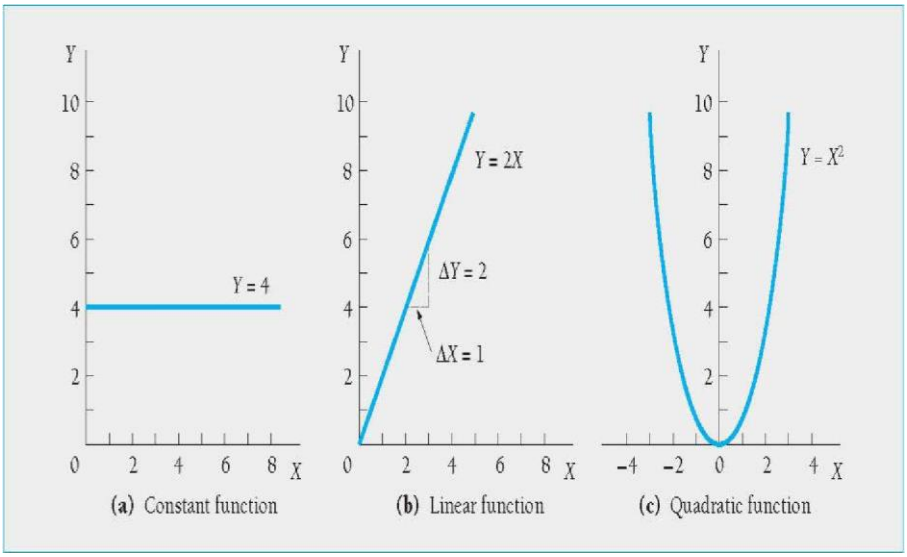


*Figure 2.3. Constant, linear and quadratic functions*

These rules for differentiating functions are summarized in Table 2.1 below.

*Table 2.1. Rules for differentiating functions*

| Function | Derivative |
|---|---|
| 1. Constant Function<br>$Y = a$ | $\frac{dY}{dX} = 0$ |
| 2. Power Function<br>$Y = aX^b$ | $\frac{dY}{dX} = b \cdot a \cdot X^{b-1}$ |
| 3. Sums of Functions<br>$Y = f_1(X) + f_2(X)$ | $\frac{dY}{dX} = \frac{df_1(X)}{dX} + \frac{df_2(X)}{dX}$ |
| 4. Product of Two Functions<br>$Y = f_1(X) \cdot f_2(X)$ | $\frac{dY}{dX} = f_1(X) \cdot \frac{df_2(X)}{dX} + f_2(X) \cdot \frac{df_1(X)}{dX}$ |
| 5. Quotient of Two Functions<br>$Y = \frac{f_1(X)}{f_2(X)}$ | $\frac{dY}{dX} = \frac{f_2(X) \cdot \frac{df_1(X)}{dX} - f_1(X) \cdot \frac{df_2(X)}{dX}}{[f_2(X)]^2}$ |
| 6. Functions of a Function<br>$Y = f_1(Z)$, where $Z = f_2(X)$ | $\frac{dY}{dX} = \frac{dY}{dZ} \cdot \frac{dZ}{dX}$ |

**Example:** Rules of differentiation: profit maximization at *Mekelle Power Co. (continued)* As an example of the application of these rules, consider again the profit function for Mekelle Power Co., given by Equation 2.6, which was discussed earlier: $\pi = -40 + 140Q - 10Q^{2.}$ In this example $Q$ represents the $X$ variable and $\pi$ represents the $Y$ variable; that is, $\pi = f(Q)$. The function $f(Q)$ is the sum of *three* separate functions—a constant function, $f_1(Q) = -40$, and two power functions, $f_2(Q) = 140Q$ and $f_3(Q) - 10Q^{2.}$ Therefore, applying the differentiation rules yields

$$\frac{d\pi}{dQ} = \frac{df_1(Q)}{dQ} + \frac{df_2(Q)}{dQ} + \frac{df_3(Q)}{dQ}$$

$$= 0 + 1* 140 * Q^{1-1} + 2*(-10) *Q^{2-1} = 140 – 20Q$$

This is the same result that was obtained earlier in Equation 2.11 by the differentiation process.

 ***Applications of Differential Calculus to Optimization Problems***

The reason for studying the process of differentiation and the rules for differentiating functions is that these methods can be used to find optimal solutions for many kinds of maximization and minimization problems in managerial economics.

 **Maximization Problem**

As you recall from the discussion of marginal analysis, a necessary (but not sufficient) condition for finding the maximum point on a curve (for example, maximum profits) is that the marginal value or slope of the curve at this point must be equal to zero. We can now express this condition within the framework of differential calculus. Because the derivative of a function measures the slope or marginal value at any given point, an equivalent necessary condition for finding the maximum value of a function $Y = f(X)$ is that the derivative $dY/dX$ at this point must be equal to zero. This is known as the *first-order condition* for locating one or more maximum or minimum points of an algebraic function.

**Example*:*** First-order condition: Profit maximization at Mekelle Power Co. (contd.)

Using the profit function (Equation 2.6)

$$\pi = -40 + 140Q - 10Q^2$$

discussed earlier, we can illustrate how to find the profit-maximizing output level $Q$ by means of this condition. Setting the first derivative of this function (which was computed previously) to zero, we obtain $d\pi/ dQ = 140 – 20Q$

$0 = 140 – 20Q$

Solving this equation for $Q$ yields $Q^* = 7$ units as the profit-maximizing output level. The profit and first derivative functions and optimal solution are shown in Figure 2.4. As we can see, profits are maximized at the point where the function is neither increasing nor decreasing; in other words, where the slope (or first derivative) is equal to zero.

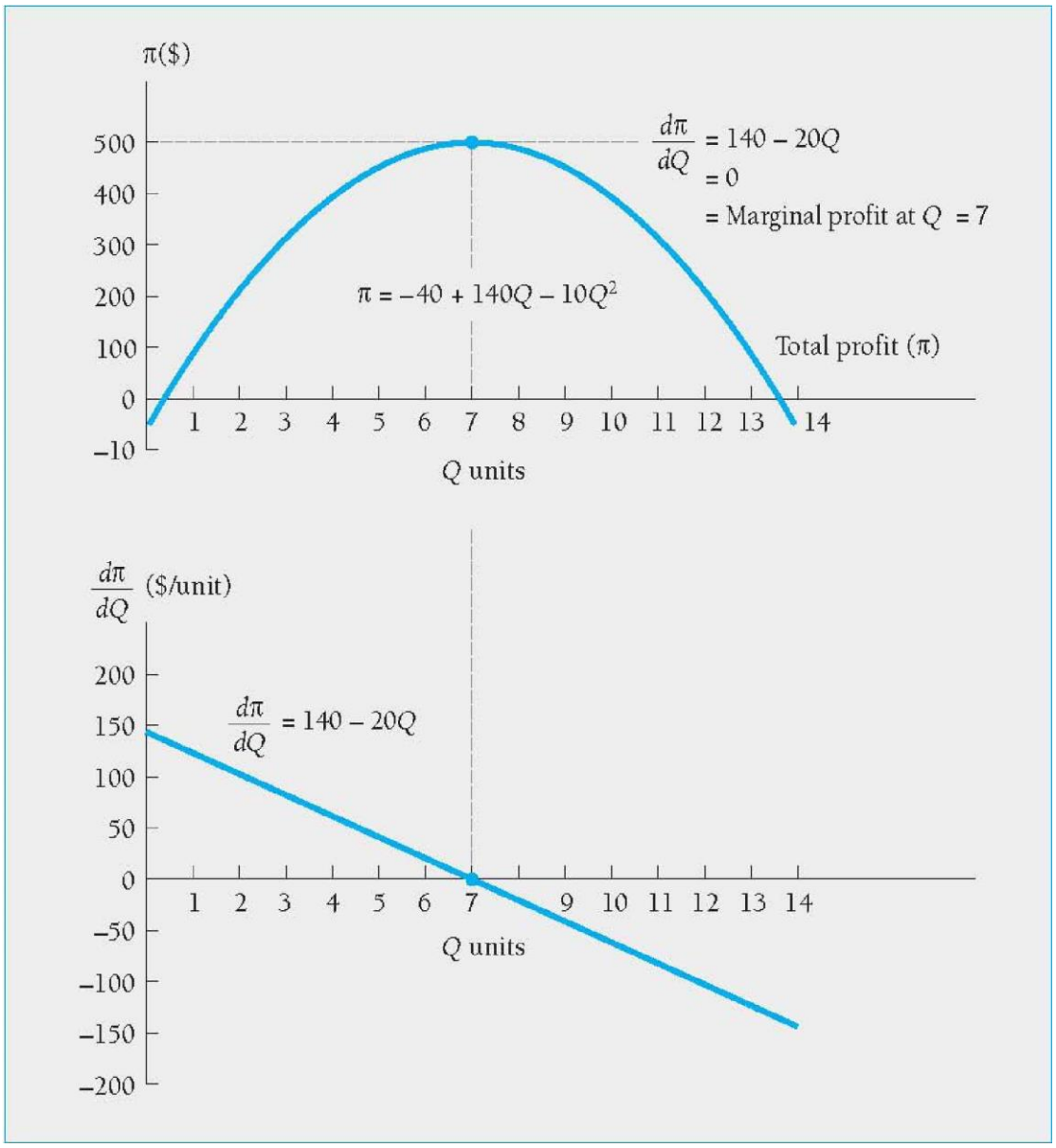


*Figure 2.4. Profit and first derivative functions*

- **Second Derivatives and the Second-Order Condition**

Setting the derivative of a function equal to zero and solving the resulting equation for the value of the decision variable does not guarantee that the point will be obtained at which the function takes on its maximum value. The slope of a U-shaped function will also be equal to zero at its low point and the function will take on its *minimum* value at the given point. In other words, setting the derivative to zero is only a *necessary* condition for finding the maximum value of a function it is not a *sufficient* condition.

Another condition, known as the *second-order condition*, is required to determine whether a point that has been determined from the first-order condition is either a maximum point or minimum point of the algebraic function. This situation is illustrated in Figure 2.5. At

both points *A* and *B*, the slope of first order condition is either a maximum point or the function (first derivative, *dY*/*dX*) is zero; however, only at point *B* does the function take on its maximum value.

We note in Figure 2.5 that the marginal value algebraic function. (slope) is continually *decreasing* in the neighbourhood of the maximum value (point *B*) of the *Y f*(*X*) function. First the slope is positive up to the point where $dY/dX = 0$, and thereafter the slope becomes negative. Thus, we must determine whether the slope 's marginal value (slope of the slope) is declining.

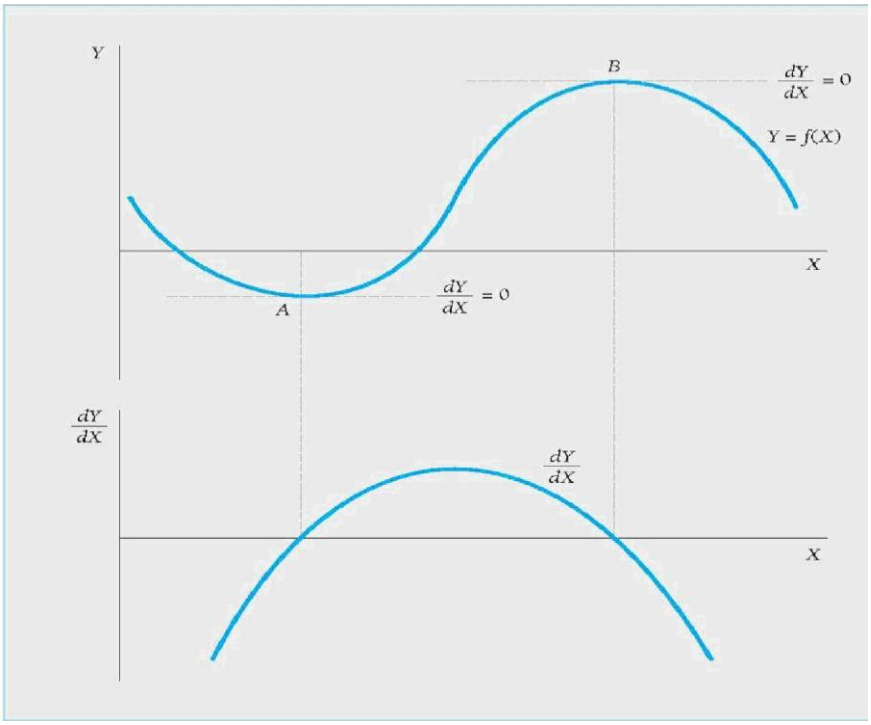


*Figure 2.5. Maximum and Minimum Values of a Function*

A test to see whether the marginal value is decreasing is to take the derivative of the marginal value and check to see if it is negative at the given point on the function. In effect, we need to find the derivative of the derivative-that is, the *second derivative* of the function-and then test to see if it is less than zero. Formally, the second derivative of the function *Y f*(*X*) is written as $d^2Y/dX^2$ and is found by applying the previously described differentiation rules to the first derivative. A *maximum point is obtained if the second derivative is negative; that is,* $d^2Y/dX^2 < 0$.

**Example:** Second-order condition: Profit maximization at Mekelle Power Co.(continued) Returning to the profit-maximization example, the second derivative is obtained from the first derivative as follows: $d\pi/dQ = 140 - 20Q$

$d^2\pi/dQ^2 = 0 + 1^* (-20)\, Q^{1-1} = -20$

Because $d^2\pi/dQ^2<0$, we know that a maximum-profit point has been obtained. An opposite condition holds for obtaining the point at which the function takes on a minimum value. Note again in Figure 2.5 that the marginal value (slope) is continually *increasing* in the neighbourhood of the minimum value (point *A*) of the *Y* (*X*) function. First the slope is negative up to the point where $dY/dX = 0$, and thereafter the slope becomes positive. Therefore, we test to see if $d^2Y/dX^2 >0$ at the given point. *A minimum point is obtained if the second derivative is positive; that is,* $d^2Y/dX^2 >0$.

- **Minimization Problem**

In some decision-making situations, cost minimization may be the objective. As in profit maximization problems, differential calculus can be used to locate the optimal points.

**Example***: Cost Minimization: MAA Garment factory*

Suppose we are interested in determining the output level that minimizes average total costs for *MAA Garment factory*, where the average total cost function might be approximated by the following relationship (*Q* represents output):

$C = 15 - .040Q + .000080Q^2$

Differentiating *C* with respect to *Q* gives

$dC/ dQ = - .040 + .000160Q$

Setting this derivative equal to zero and solving for *Q* yields

$0 = .040 + .000160Q$

$Q^* = 250$

Taking the second derivative, we obtain

$d^2C/ dQ^2 = + .000160$

Because the second derivative is positive, the output level of $Q = 250$ is indeed the value that minimizes average total costs.

### 2.2.2. Partial Differentiation and Multi-variate Optimization

Until now we have examined the relationship between two variables only. For example, variable *Y* (say, total revenue, total cost, or total profit) was assumed to be a function of or to depend on only the value of variable *X* (total output or quantity). Most economic relationships, however, involve more than two variables. For example, total revenue may be a function of or depend on both output and advertising, total costs may depend on expenditures on both labor and capital, and total profit on sales of commodities *X* and *Y*. Thus, it becomes important to determine the marginal effect on the dependent variable, say,

total profit, resulting from changes in the quantities of each individual variable, say, the quantity sold of commodity *X* and commodity *Y, separately.* These marginal effects are measured by the *partial derivative,* which is indicated by the symbol $\partial$ (as compared to *d* for the derivative). The partial derivative of the dependent or left-hand variable with respect to each of the independent or right-hand variables is found by the same rules of differentiation presented earlier, except that all independent variables other than the one with respect to which we are finding the partial derivative are held constant.

For example, suppose that the total profit ($\pi$) function of a firm depends on sales of commodities *X* and *Y* as follows:

$\pi = f(X, Y) = 80X - 2X^2 - XY - 3Y^2 + 100Y$ .....*Equation 2.12*

To find the partial derivative of $\pi$ with respect to *X*, $\partial\pi/\partial X$, we hold *Y* constant and obtain

$\pi = 80 - 4XY - Y$

This isolates the marginal effect on $\pi$ from changes in the quantity sold of commodity *X* only (i.e., while holding the quantity of commodity *Y* constant). Note that the derivative of the third term of the $\pi$ function is $-Y$ (since the implicit exponent of *X* is 1) and that *Y* is treated as a constant. The fourth and the fifth terms of the $\pi$ function drop out in the partial differentiation because they contain no *X* term. Similarly, to isolate the marginal effect of a change of *Y* on $\pi$, we hold *X* constant and obtain $\pi = -X - 6Y + 100$

We can visualize geometrically the concept of the partial derivative with a three-dimensional figure, with $\pi$ on the vertical axis and with the *X* axis and the *Y* axis forming the (plane surface, rather than the line) base of the figure. Then, $\partial\pi/\partial X$ measures the marginal effect of

*X* on $\pi$, in the cross section of the three-dimensional figure along the *X* axis. Similarly, $\partial\pi/\partial Y$ examines the marginal effect of *Y* on $\pi$ in the cross section of the three-dimensional figure along the *Y* axis. Note also that the value of $\partial\pi/\partial X$ depends also on the level at which *Y* is held constant. Similarly, the value of $\partial\pi/\partial Y$ depends also on the level at which *X* is held constant. This is the reason that the expression for the $\partial\pi/\partial X$ *found* above also contains a *Y* term, while $\partial\pi/\partial Y$ also has an *X* term.

- *Maximizing a Multivariable Function*

To maximize or minimize a multivariable function, we must set each partial derivative equal to zero and solve the resulting set of simultaneous equations for the optimal value of the independent or right-hand variables. For example, to maximize the total-profit function

(Equation 2.12, repeated below for ease of reference),

$\pi = 80X - 2X2 - XY - 3Y2 + 100Y$……..……. *Equation 2.12*

We set $\partial\pi/\partial X$ and $\partial\pi/\partial Y$ (found earlier) equal to zero and solve for $X$ and $Y$.

Specifically,

$\partial\pi/\partial X = 80 - 4X - Y = 0$

$\partial\pi/\partial Y = - X - 6Y + 100 = 0$

Multiplying the first of the above expressions by -6, rearranging the second, and adding, we get

$-480 + 24X + 6Y = 0$

$\underline{100 - X - 6Y = 0}$

$-380 + 23X = 0$

Therefore, $X = 380/23 = 16.52$. Substituting $X = 16.52$ into the first expression of the partial derivative set equal to zero, and solving for $Y$, we get

$80 - 4(16.52) - Y = 0$

Therefore, $Y = 80 - 66.08 = 13.92$.

Thus, the firm maximizes $\pi$ when it sells 16.52 units of commodity $X$ and 13.92 units of commodity $Y$. Substituting these values into the $\pi$ function, we get the maximum total profit of the firm of $\pi = 80(16.52) - 2(16.52)^2 - (16.52)(13.92) - 3(13.92)^2 + 100(13.92) =$ **$1,356.52**

### 2.2.3. Constrained Optimization and Lagrangian Multiplier Techniques

Most organizations have constraints on their decision variables. The most obvious constraints, and the easiest to quantify and incorporate into the analysis, are the limitations imposed by the quantities of resources (such as capital, personnel, facilities, and raw materials) available to the organization. Other more subjective constraints include legal, environmental, and behavioural limitations on the decisions of the organization.

- ***Substitution Technique***

When the constraints take the form of equality relationships, classical optimization procedures can be used to solve the problem. Substitution method can be employed when the objective function is subject to only *one* constraint equation of a relatively simple form, is to solve the constraint equation for one of the decision variables and then substitute this expression into the objective function. This procedure converts the original problem into an unconstrained optimization problem, which can be solved using the calculus procedures

developed. This procedure can be clarified by examining its use in a constrained minimization problem.

Suppose a firm produces its product on two assembly lines and operates with the following total cost function:

$TC = \$3X^2 + \$6Y^2 - \$1XY$ where $X$ represents the output produced on one assembly line and $Y$ the production from the second.

Management seeks to determine the least-cost combination of $X$ and $Y$, subject to the constraint that total output of the product is 20 units. The constrained optimization problem is Minimize $TC = \$3X^2 + \$6Y^2 - \$1XY$ subject to; $X + Y = 20$

Solving the constraint for $X$ and substituting this value into the objective function results in

$X = 20 - Y$ and

$TC = \$3(20 - Y)^2 + \$6Y^2 - \$1(20 - Y)\ Y$

$= \$3(400 - 40Y + Y^2) + \$6Y^2 - \$1(20Y - Y^2)$ ……. *Equation 2.13*

$= \$1{,}200 - \$120Y + \$3Y^2 + \$6Y^2 - \$20Y + Y^2$

$= \$1{,}200 - \$140Y + \$10Y^2$

Now it is possible to treat Equation 2.13 as an unconstrained minimization problem. Solving it requires taking the derivative of the total cost function, setting that derivative equal to zero, and solving for the value of $Y$: $dTC/\ dY = -\$140 + \$20Y = 0$

$20Y = 140$

$Y = 7$

A check of the sign of the second derivative evaluated at that point ensures that a minimum has been located: $dTC/\ dY = -\$140 + \$20Y$ $d^2\ TC/\ dY^2 = 20$

Because the second derivative is positive, $Y = 7$ is indeed a minimum. Substituting 7 for $Y$ in the constraint equation allows one to determine the optimal quantity to be produced on assembly line $X$:

$X + 7 = 20$

$X = 13$

Thus, production of 13 units of output on assembly line $X$ and seven units on line $Y$ is the least-cost combination for manufacturing a total of 20 units of the firm ‘s product. The total cost of producing that combination is

$TC = \$3(132) + \$6(72) - \$1(13 \times 7)$

= $507 + $294 –

$91 = $710

 ***Lagrangian technique***

Unfortunately, the substitution technique used in the preceding section is not always feasible. Constraint conditions are sometimes too numerous or complex for substitution to be used. In these cases, the technique of *Lagrangian multipliers* can be used.

The Lagrangian technique for solving constrained optimization problems is a method that calls for optimizing a function that incorporates the original objective function and the constraint conditions. This combined equation, called the Lagrangian function, is created in such a way that when it is maximized or minimized the original objective function is also maximized or minimized, and all constraints are satisfied.

A re-examination of the constrained minimization problem shown previously illustrates this technique. Recall that the firm sought to minimize the function

$TC = \$3X^2 + \$6Y^2 - \$1XY$, subject to the constraint that $X + Y = 20$. Rearrange the constraint to bring all terms to the right of the equal sign: $0 = 20 - X - Y$

This is always the first step in forming a Lagrangian expression. Multiplying this form of the constraint by the unknown factor λ and adding the result to the original objective function creates the Lagrangian expression:

$LTC = \$3X^2 + \$6Y^2 - \$1XY + \lambda (20 - X - Y)$ ……. *Equation 2.14*

*LTC* is defined as the Lagrangian function for the constrained optimization problem under consideration. Because it incorporates the constraint into the objective function, the Lagrangian function can be treated as an unconstrained optimization problem. The solution to the unconstrained Lagrangian problem is *always* identical to the solution of the original constrained optimization problem.

To illustrate, consider the problem of minimizing the Lagrangian function constructed in Equation 2.14. At a minimum point on a multivariate function, all partial derivatives must equal zero. The partials of Equation 2.14 can be taken with respect to the three unknown variables, *X*, *Y*, and λ, as follows: $\partial LTC/\partial X = 6X - Y - \lambda$

$\partial LTC/\partial Y = 12Y - X - \lambda$ and

$\partial LTC/\partial \lambda = 20 - X - Y$

Setting these three partials equal to zero results in a system of three equations and three unknowns:

$6X - Y - \lambda = 0$……...…… *Equation 2.15*

$-X + 12Y - \lambda = 0$……..……...*Equation 2.16*

$20 - X - Y = 0$……..……*Equation 2.17*

Notice that Equation 2.17, the partial of the Lagrangian function with respect to λ, is the constraint condition imposed on the original optimization problem. The Lagrangian function is constructed so that the derivative of the function taken with respect to the Lagrangian multiplier, λ, always gives the original constraint. So long as this derivative is zero, as it must be at a local extreme (maximum or minimum), the constraint conditions imposed on the original problem are met.

Further, because the last term in the Lagrangian expression must equal zero ($0 = 20 - X - Y$), the Lagrangian function reduces to the original objective function, and the solution to the unconstrained Lagrangian problem is always the solution to the original constrained optimization problem. Completing the analysis for the example illuminates these relations.

To begin, it is necessary to solve the system of equations to obtain optimal values of *X* and *Y*.

Subtracting Equation 2.16 from Equation 2.15 gives

$7X - 13Y = 0$……...……*Equation 2.18*

Multiplying Equation 2.7 by 7 and adding Equation 2.18 to this product gives the solution for *Y*:

$$\left.\begin{aligned} 140 - 7X - 7Y &= 0 \\ 7X - 13Y &= 0 \end{aligned}\right\}$$

$140 - 20Y = 0$

$140 = 20Y$

$Y = 7$

Substituting 7 for *Y* in Equation 2.17 yields $X = 13$, the value of *X* at the point where the Lagrangian function is minimized. Because the solution of the Lagrangian function is also the solution to the firm ‘s constrained optimization problem, 13 units from assembly line *X* and seven units from line *Y* is the least cost combination of output that can be produced subject to the constraint that total output must be 20 units. This is the same answer obtained previously, using the substitution method. The Lagrangian technique is a more powerful technique for solving constrained optimization problems than the substitution method; it is easier to apply with multiple constraints, and it provides valuable supplementary

information. This is because the Lagrangian multiplier itself has an important economic interpretation.

Substituting the values of $X$ and $Y$ into Equation 2.15 gives the value of λ:

$6 \times 13 - 7 - \lambda =$

$0\ \lambda = \$71$

Here, λ is interpreted as the marginal cost of production at 20 units of output. It means that if the firm were allowed to produce only 19 instead of 20 units of output, total costs would fall by approximately \$71. If the output requirement were 21 instead of 20 units, costs would increase by roughly that amount. Because λ = \$71 can be interpreted as the marginal cost of production, an offer to purchase another unit of output for \$100 is acceptable because it results in a \$29 marginal profit. Conversely, an offer to purchase an additional unit for \$50 would be rejected because a marginal loss of \$21 would be incurred. λ can be thought of as a planning variable, because it provides valuable information concerning the effects of altering current activity levels.

Another example provides additional perspective on the Lagrangian method. Recall from the discussion of Equation 2.16 and Figure 2.5 that the profit function, $\pi = -\$10{,}000 + \$400Q - \$2Q^2$ where π is total profit and $Q$ is output in units, is maximized at $Q = 100$ with $\pi = \$10{,}000$. The impact of constraints in the production process, and the value of the Lagrangian method, can be portrayed by considering the situation in which each unit of output requires 4 hours of skilled labour, and a total of only 300 hours of skilled labour is currently available to the firm. In this instance, the firm seeks to maximize the function $\pi = -\$10{,}000 + \$400Q - \$2Q^2$, subject to the constraint $4Q = 300$ (because $L = 4Q$). Rearrange the constraint to bring all terms to the right of the equal sign:

$0 = 300 - 4Q$

Multiplying this form of the constraint by λ and adding the result to the original objective function creates the Lagrangian expression:

$L\pi = -\$10{,}000 + \$400Q - \$2Q^2 + \lambda\,(300 - 4Q)$ ……...……. *Equation 2.19*

with the following partials: $\partial L\pi/\partial Q = 400 - 4Q - 4\,\lambda$ *and* $\partial L\pi/\partial\,\lambda = 300 - 4Q$

Setting these two partials equal to zero results in a system of two equations and two unknowns. Solving provides the values $Q = 75$, $\lambda = \$25$, and, from the objective function, $\pi = \$8{,}750$. The constraint on skilled labour has reduced output from 100 to 75 units and has reduced total profits from \$10,000 to \$8,750. The value λ = \$25 indicates that should a one-

unit expansion in output become possible, total profits would rise by $25. This information indicates that the maximum value of additional skilled labour is $6.25 per hour, because each unit of output requires 4 hours of labour. Assuming there are no other costs involved, $6.25 per hour is the most the firm would pay to expand employment.

The effects of relaxing the constraint as progressively more skilled labour becomes available are illustrated in Figure 2.5. If an additional 100 hours of skilled labour, or 400 hours in total, is available, the output constraint would become $0 = 400 - 4Q$, and solved values $Q = 100$, $\lambda = \$0$, and $\pi = \$10{,}000$ would result. The value $\lambda = \$0$ indicates that skilled labour no longer constrains profits when 400 hours are available. Profits are maximized at $Q = 100$, which is the same result obtained in the earlier unconstrained solution to this profit maximization problem.

In this instance, the output constraint becomes nonbinding because it does not limit the profitmaking ability of the firm. Indeed, the firm is not willing to employ more than 400 hours of skilled labour. To illustrate this point, consider the use of 500 hours of skilled labour and the resulting constraint $0 = 500 - 4Q$. Solved values are $Q = 125$, $\lambda = -\$25$, and $\pi = \$8{,}750$.

The value $\lambda = -\$25$ indicates that one additional unit of output, and the expansion in employment that results, would reduce profits by $25. Conversely, a one-unit reduction in the level of output would increase profits by $25. Clearly, the situation in which $\lambda < 0$ gives the firm an incentive to reduce input usage and output, just as $\lambda > 0$ provides an incentive for growth.

To generalize, a Lagrangian multiplier, $\lambda$, indicates the marginal effect on the objective function of decreasing or increasing the constraint requirement by one unit. Often, as in the previous examples, the marginal relation described by the Lagrangian multiplier provides economic data that help managers evaluate the potential benefits or costs of relaxing constraints. Thus, the Lagrangian multiplier technique creates an additional artificial variable for each constraint. Using these artificial variables, the constraints are incorporated into the objective function in such a way as to leave the value of the function unchanged. If a problem has two or more constraints, then a separate variable is defined for each constraint and incorporated into the Lagrangian function. In general, $\lambda$ measures the marginal change in the value of the objective function resulting from a one-unit change in the value on the right-hand side of the equality sign in the constraint relationship.

***Self Assessment Exercises***

1.Why is the first-order condition for finding a maximum (or minimum) of a function referred to as a necessary but not sufficient condition?

2.Defining $Q$ to be the level of output produced and sold, assume that the firm 's cost function is given by the relationship $TC = 20 + 5Q + Q^2$

Furthermore, assume that the demand for the output of the firm is a function of price $P$ given by the relationship $Q = 25 - P$

a) Defining total profit as the difference between total revenue and total cost, express in terms of $Q$ the total profit function for the firm. (*Note:* Total revenue equals price per unit times the number of units sold.)

b) Determine the output level where total profits are maximized.

c)Calculate total profits and selling price at the profit-maximizing output level.

d)If fixed costs increase from $20 to $25 in the total cost relationship, determine the effects of such an increase on the profit-maximizing output level and total profits.

3.Suppose that a firm 's profit function is given by the following relationship: $\pi = -25 + 100Q_1 + 95Q_2 - 10Q_1^2 - 5Q_2^2 - 5Q_1Q_2$

Where $Q_1$ and $Q_2$ are the respective quantities of the two products that the firm manufactures and sells. Determine:

a) the values of $Q_1$ and $Q_2$ that maximize $\pi$

b) the maximum value of $\pi$

4.Unilever Ethiopia PLC. has determined through regression analysis that its sales ($S$) are a function of the amount of advertising (measured in units) in two different media.

This is given by the following relationship ($X$ = newspapers, $Y$ = magazines):

$S(X, Y) = 200X + 100Y - 10X^2 - 20Y^2 + 20XY$

f. Find the level of newspaper and magazine advertising that maximizes the firm 's sales.

g. Calculate the firm 's sales at the optimal values of newspaper and magazine advertising determined in part (a).

5.Consider Problem 4 again. Suppose that the advertising budget is restricted to 20 units.

a. Determine (using Lagrangian multiplier techniques) the level of newspaper and magazine advertising that maximizes sales subject to this budget constraint.

b. Calculate the firm 's sales at this constrained optimal level.

# CHAPTER THREE
# 3. PRODUCTION AND COST ANALYSIS

***Objectives***

- *To introduce the concept of production and explain its relevance to managerial decision-making.*
- *To describe and explain relationships between total, average and marginal product, and the different stages of production.*
- *To explain the theory of cost and the parallels between production theory and cost theory.*
- *To explain the meaning of demand estimation and examine different methods of demand estimation*

## 3.1. Basic Concepts of Production theory

The firm is an organizational activity that transforms factors of production, or productive inputs, into outputs of goods and services. Economics itself is the study of how society chooses to satisfy virtually unlimited human wants subject to scarce productive resources. In fact, this economic problem can be viewed as a constrained optimization problem in which the objective is to maximize some index of human happiness, which is assumed to be a function of the consumption of limited, or scarce, amounts of goods and services.

Let us examine this fundamental economic paradigm a bit more closely. For theoretical purposes, it is useful to assume that human material wants and desires are insatiable. While this might not be so for any individual commodity, by and large the vast majority of individuals are never completely satisfied with what they have. Consumers will, in general, always want more of something. Unfortunately, consumers cannot have everything they want for the simple reason that the productive inputs necessary to produce these goods and services are finite. In other words, outputs of goods and services (*Q*) are limited because the productive resources necessary to produce them are limited.

*Productive resources*, or *inputs*, or *factors of production* are used by firms to produce goods and services. These productive resources may conceptually be divided into two broad categories-human and nonhuman resources. Nonhuman resources may be further classified as *land*, *raw materials*, and *capital*. While these classifications are arbitrary, they are conceptually convenient. Land (e.g., arable land for farming, or the land under buildings and roads) may be described as a ―gift of nature. Raw materials include such natural

resources as coal, oil, copper, sand, and trees. Finally, capital represents manufactured inputs that are used in the production of final goods and services. Capital goods include tools, machinery, equipment, office buildings, and storage facilities.

In an economic sense, ―capital‖ does not refer to such financial assets as money, corporate equities, savings accounts, and Treasury securities. There is a loose conceptual connection between the economic and financial definitions of the term in that firms that wish to acquire capital equipment will do so only if the additional return from an investment in capital goods is sufficient to cover the opportunity cost of the interest that might have been earned by investing in a financial asset. Human resources, on the other hand, might be classified as *labour* and *entrepreneurial ability*. Labour consists of the physical and mental talents of individuals engaged in the production process. Labour services are similar to services derived from nonhuman resources in that the rental value of labour services (wages, salaries, health benefits, etc.) is determined by the interaction of supply and demand conditions in the market for factors of production.

Labour is unique among the other factors of production, however, in that in a free society, labour decides for itself how intensively its services will be made available in the production process. Entrepreneurial ability is a special subset of human resources because without the entrepreneur, all other factors of production cannot be combined. In other words, the services of land, labour, and capital are ―rented‖ to the highest bidder through the interaction of supply and demand. The bidders are entrepreneurs. An individual may, of course, be both a labourer and entrepreneur in the same production process. Nevertheless, the entrepreneur is unique in that it is through his or her initiative and enterprise that the other factors of production are organized in the first place. The entrepreneur is distinguished from the ―hired hand‖ in a number of other ways. The entrepreneur, for example, makes the non-routine decisions for the firm. The entrepreneur determines the firm ‘s organizational objectives. The entrepreneur is an innovator, constantly on the alert for the best (least expensive) production techniques. The entrepreneur also endeavours to ascertain the wants and needs of the public to be able to introduce new and better products, and to discontinue production of goods and services that are no longer in demand. The entrepreneur is a risk taker in the expectation of earning a profit.

It should be pointed out once again that it is assumed that the expectation of earning profit is the motivating incentive for all commercial activities in a free economy. Without the

profit motive, it is assumed that entrepreneurs would not incur risk the financial risks associated with combining productive resources to produce final goods and services for sale in the market. In the absence of the profit motive, the production of goods and services in quantities that are most in demand by society would not occur.

Moreover, whenever the profit incentive is diluted-saying, through a tax on profits-fewer goods and services will be produced by the private sector in exchange for an increase in goods and services produced by the public sector. Whether society is better off or worse off will ultimately depend on which sector is most responsive to society ‘s needs and preferences and which sector is most efficient in the utilization of productive resources in the delivery of those goods and services. The disappointing economic experiences of centrally planned economies in the communist countries of eastern Europe and Asia in the latter half of the twentieth century, compared with that of the essentially free market economies of western Europe, North America, and Japan often convincing testimony to the power of the profit motive in promoting the general economic well-being of society at large.

To quote from

Adam Smith: ―By pursuing his own interest he (the entrepreneur) frequently promotes that of society more effectively than when he really intends to promote it.

### 3.1.1. The Production Function

The technological relationship that describes the process whereby factors of production are efficiently transformed into goods and services is called the *production function*. Mathematically, a production function utilizing capital, labour, and land inputs may be written as

$Q = f(K, L, M)$ .....*Equation 3.1*

Where $K$ is capital, $L$ is labour, and $M$ is land

The production function defines the maximum rate of output per unit of time obtainable from a given set of productive inputs. Obviously, some firms organize inputs inefficiently, thereby producing less than the maximum possible level of goods and services. In a competitive market environment, however, firms that do not adopt the most efficient (cost-effective) production technology will not prosper and may, if fact, be forced to curtail operations as their market position is undermined by more aggressive competitors. In free-market economies there is a tendency for more efficient operations to drive less efficient firms out of an industry. For the sake of pedagogical, theoretical, and graphical convenience,

let us assume that all productive inputs may be classified as either labour ($L$) or capital goods ($K$). Equation (3.1) may therefore be rewritten as

$Q = f(K, L)$

In most cases, labour and capital may be substituted, albeit in varying degrees. As we will see, the precise manner in which these inputs are combined will depend on the relative rental price and marginal productivity of the factor of production. A firm that operates efficiently will choose that combination of productive inputs that minimizes the total cost of producing a given level of output. Equivalently, a firm that operates efficiently will choose that combination of productive inputs that will maximize production subject to a given operating budget. It should be noted that the production function signifies the technological relationship between inputs and outputs. It is an engineering concept, devoid of any economic content. The production function defines the maximum output obtainable from a particular combination of inputs. Input prices must be combined with the production function to determine which of the many possible combinations of inputs is the most desirable given the firm ‘s objectives, such as profit maximization.

 **Short – run production function**

In the production process, economists distinguish between the *short run* and the *long run*. The short run in production refers to that period of time during which at least one factor of production is held constant. The law of diminishing marginal product is a short-run phenomenon. The long run in production occurs when all factors of production are variable. Denoting $K_0$ as a fixed amount of capital, the short-run production function may be written

$QL = f(K_0, L) = TPL$.....*Equation3.2*

This Equation is sometimes referred to as to *total product of labour*, or more simply as *TPL*. The decision maker must determine the optimal amount of the variable input ($L$) given its price ($PL$) and the price of the resulting output ($P$). It is assumed that the state of technology is captured in the specification of the production function and that production is efficient. Note that the subscript $L$ in Equation (3.2) denotes the level of output dependent on the level of labour usage. Equivalently, if labour is assumed to be the fixed input, the short run production function would be written

$QK = f(K, L_0) = TPK$.....*Equation 3.3*

### 3.1.2. Key relationships: total, average, and marginal products

There is a crucial relationship between any total concept and its related average and marginal concepts. Since the present discussion deals with production, we will explore the

relationship between the total product of labour and the related *average product of labour* (*APL*) and *marginal product of labour* (*MPL*). The reader is advised, however, that the fundamental relationships established here are perfectly general and may be applied to any functional relationship.

- **Total product of Labour (TPL)**

If we assume that the level of capital usage is fixed, the total product of labour is described by Equation (3.2). This short-run production function defines the maximum amount of output attainable from any given level of labour usage given some fixed level of capital usage.

- **Average product of Labour (APL)**

The average product of labour, *APL*, is simply the total product per unit of labour usage and is determined by dividing the total product of labour by the total amount of labour usage. This relationship is defined in Equation (3.4).

$$AP_L = \frac{TP_L}{L} = \frac{Q_L}{L} = \frac{f(K_0, L)}{L}$$ .....*Equation 3.4*

- **Marginal Product of Labour (MPL)**

The *marginal product of labour MPL* is defined as the incremental change in output associated with an incremental change in the amount of labour usage. Mathematically, since the amount of capital is constant, this is equivalent to the first partial derivative of the production function with respect to labour:

$$MP_L = \frac{\partial TP_L}{\partial L} = \frac{\partial Q_L}{\partial L} = \frac{\partial f(K_0, L)}{\partial L}$$ .....*Equation 3.5*

Similarly, the marginal product of capital is defined as

$$MP_K = \frac{\partial TP_K}{\partial K} = \frac{\partial Q_K}{\partial K} = \frac{\partial f(K, L_0)}{\partial K}$$ .....*Equation 3.6*

### 3.1.3. Mathematical relationship between APL and MPL

The mathematical relationship between the average product of labour (or any average concept) and the marginal product of labour (or any related marginal concept) may be illustrated by the use of optimization analysis. Consider again the definition of the average product of labour

$$AP_L = \frac{TP_L}{L} = \frac{Q_L}{L} = \frac{f(K_0, L)}{L}$$ .....*Equation 3.7*

Taking the first derivative of Equation (3.7) with respect to labour and setting the results equal to zero yields

$$\frac{\partial AP_L}{\partial L} = \frac{L(\partial Q_L/\partial L) - Q_L(\partial L/\partial L)}{L^2} = 0 \quad \text{.....Equation 3.8}$$

Which implies

$$MP_L - \frac{Q_L}{L} = 0 \quad \text{or} \quad MP_L = AP_L \quad \text{.....Equation 3.9}$$

Since $1/L^2$ is positive. That is, when the average product of labour is maximized, the marginal product of labour is equal to the average product of labour. Equation (3.9) may also be used to highlight another important relationship between the average product of labour and the marginal product of labour. When $MPL > APL$, then $\partial APL/\partial L > 0$, which implies that the average product of labour is rising. When $MPL < APL$, then $\partial APL/\partial L < 0$, which says that the average product of labour is falling. Only when $MPL = APL$ and $\partial APL/\partial L = 0$ will the average product of labour be stationary (either a maximum or a minimum). The second-order condition for minimum average product of labour is $\partial^2 APL/\partial L^2 < 0$.

In general, for any total function it can be easily demonstrated that when the marginal value is greater than the average value (i.e., $M > A$), then $A$ will be rising. When $M < A$, $A$ will be falling. Finally, when $M = A$, $A$ will neither be falling nor rising but at a local optimum (maximum or minimum). To illustrate this relationship, consider the simple example of a student who takes a course in which the final grade is based on the average of 10 quizzes. A maximum of 100 points may be earned on each quiz. Thus, a maximum of 1,000 points may be earned during the semester, or a maximum average for the course of 1,000/10 = 100. Suppose that the student has taken six quizzes and has earned a total of 480 points, for an average grade of 480/6 = 80. If the student receives a grade of 90 on the seventh quiz, then the student ‘s average will rise from 80 to 570/7 = 81.4. That is, since the marginal grade is greater than the average grade to that point, the student ‘s average will rise. On the other hand, if the student received a grade of 70 on the seventh quiz, the student ‘s average will fall to 550/7 =78.6. Finally, if the student receives a grade of 80 on the seventh quiz then, clearly, there will be no change in the student ‘s average (i.e., 560/7 = 80).

### 3.1.4. The Law of Diminishing Marginal Product

The Cobb–Douglas production function exhibits a number of useful mathematical properties. One of these properties is the important technological relationship known as the

*law of diminishing marginal product* (*law of diminishing returns*). This concept can be described with the use of a simple illustration. Consider a tomato farmer who has a 10-acre farm and as much fertilizer, capital equipment, water, labour, and other productive resources as is necessary to grow tomatoes. The only input that is fixed in supply is farm acreage. The farmer decides that to maximize output, additional workers will have to be hired. With the exception of farm acreage, each worker has as many productive resources to work with as necessary. Initially, as one might expect, output expands rapidly. At least in the early stages of production, as more workers are assigned to the cultivation of tomatoes, the additional output per worker might be expected to increase. This is because in the beginning land is relatively abundant and labour is relatively scarce. While each worker has as much land and other resources to work with as is necessary for efficient production, at least some land initially stands fallow. Labour can be said to be fully utilized while land can be said to be underutilized. As more labourers are added to the production process, total output rises; beyond some level of labour usage, however, incremental additions to output from the addition of more workers, while positive, will begin to decline. That is, while each additional worker contributes positively to total output, beyond some point the amount of land allocated to each worker will decline. No matter how much water, fertilizer, and other inputs are made available to each worker, the amount of output per worker will begin to fall. At this point, land has become over utilized while labour has become fully utilized.

The law of diminishing marginal product sets in at the point at which the contribution to total output from an additional worker begins to fall. In fact, if successively more workers are added to the production process, the amount of land allocated to each worker becomes so small that we might even expect zero marginal product; that is, total output has been maximized. It is even conceivable that beyond the point of maximum production, as more workers are added to the production process, output will actually decline. This is because workers may interfere with each other or will perhaps trample on some of the tomatoes. In the extreme, we cannot rule out the possibility of negative marginal product of a variable input.

*Definition*: The law of diminishing marginal product states that as increasing amounts of a variable input are combined with one or more fixed inputs, at some point the marginal product of the variable input will begin to decline.

 **Relationships among the production function**

Figure 3.1 below illustrates the short-run relationships among the total, average, and marginal product functions, which the marginal product of labour, which is the first partial derivative of the total product of labour function, is the value of the slope of a tangent at a particular point on the *TPL* curve .We can see from Figure 3.1.a, that as we move from point 0 on the *TPL* curve, the slope of the tangent steadily increases in value until we reach inflection point *A*. From point *A* to point *C* the marginal product of labour is still positive, but its value steadily decreases to zero, which is at the top of the *TPL* ―hill.‖ Beyond point *C* the slope of the value of the tangent (*MPL*) becomes negative. These relationships are illustrated in Figure 3.1.b, where the *MPL* function reaches its maximum at the labour usage level 0*D*, which is at the inflection point, thereafter declining steadily until *MPL* = 0 at 0*F*. For labour usage beyond point 0*F*, then *MPL* becomes negative, resulting in the decline in *TPL*. In Figure 3.1a the distance 0*A* along *TPL* represents *increasing marginal product of labour*. This range of labor usage is represented by the distance 0*D* and is characterized by increasing incremental contributions to total output arising from incremental applications of labour input. This phenomenon, if it occurs at all, is likely to take place at lower output levels. This is because at low output levels the fixed input is likely to be underutilized, making it likely that additional applications of the variable input will result in increased efficiency arising from specialization, management, communications, and so on. Once efficiency gains from specialization have been exhausted, however, the production process will be characterized by the law of diminishing marginal product. This phenomenon sets in at the labour usage *DF*, which corresponds to the distance *AC* along the *TPL* curve. The law of diminishing marginal product sets in at the inflection point *A* and continues until *MPL* = 0 at point *C*. To reiterate, the law of diminishing marginal product says that as more and more of the variable input is added to the production process with the amount utilized of at least one factor of production remaining constant, beyond some point, incremental additions to output will become smaller and smaller. Finally, movement along the *TPL* curve beyond point *C* is characterized by *negative marginal product of labour*. Beyond the level of labour usage 0*F* in Figure 3.1a, incremental increases in labour usage will actually result in a fall in total output. In Figure 3.1b this phenomenon is illustrated by *MPL* < 0; that is, the *MPL* curve falls below the horizontal axis for output levels in excess of 0*F*.

- **Average product**

The average product of labour may be illustrated diagrammatically as the value of the slope of a ray through the origin to a given point on the total product curve. To see this, consider the definition of the marginal product of labour for discrete changes in output.

$$MP_L = \frac{\Delta Q}{\Delta L} = \frac{Q_2 - Q_1}{L_2 - L_1}$$ .....*Equation 3.10*

Suppose that we arbitrarily select as our initial price quantity combination the origin (i.e., $Q_1 = L_1 = 0$). Substituting these values into Equation (3.10) we get

$$MP_L = \frac{Q_2 - 0}{L_2 - 0} = \frac{Q_2}{L_2} = AP_L$$ .....*Equation 3.11*

which is the same result as Equation 3.9

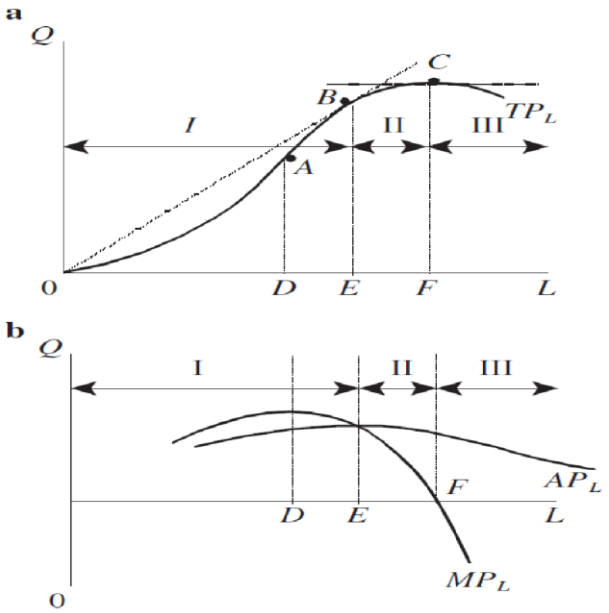


*Figure 3.1. Stages of production*

Referring to Figure 3.1a, we see that as we move up along the *TPL* curve from the origin, *APL* reaches a maximum at point *B*, steadily declining thereafter. Note also that at point *B* the average product of labour is precisely the value of the marginal product of labour because at that point the ray from the origin and the tangent at that point are identical. This is seen in Figure 3.1b at the labour usage 0*E*, where the *APL* curve intersects the *MPL* curve. Finally, unlike the marginal product of labour (or any other productive input), the average product of labour cannot be negative because labour and output can never be negative.

### 3.1.5. The three stages of production function

Figure 3.1 can also be used to define the three stages of production. *Stage I of production* is defined as the range of output from $L = 0$ to, but not including, the level of labour usage at which $APL = MPL$. Alternatively, stage I of production is defined up to the level of labour usage at which the average product of labour is maximized. In this range, labour is over utilized, whereas capital is underutilized. This can be seen by the fact that $MPL > APL$ thus ―pulling up‖ output per unit of labour. If we assume that the wage rate per worker and the price per unit of output are constant, then increasing output per worker suggests that average revenue generated per worker is rising, which suggests that average profit per worker is also rising. It stands to reason, therefore, that no firm would ever actually operate within this region of labour usage (0 to $MPL = APL$), since additions to the labour force will increase average worker productivity and, under the appropriate assumptions, average profit generated per worker as well.

*Stage II of production* is defined in Figure 3.1 as the labour usage levels $0E$ to $0F$. In this region, the marginal product of labour is positive but is less than the average product of labour, thus ―pulling down‖ output per worker, which implies that average revenue generated per worker is also falling. In this region, labour becomes increasingly less productive on average.

Finally, *stage III of production* is defined along the *TPL* function for labour input usage in excess of $0F$, where $MPL < 0$. As it is apparent that production will not take place in stage I of production because an incremental increase in labour usage will result in an increase in output per worker and, under the appropriate assumptions, an increase in profit per worker, so it is also obvious that production will not take place in stage III. This is because an increase in labour usage will result in a decline in total output accompanied by an increase in total cost of production, implying a decline in profit.

Stage III is also the counterpart to stage I of production. Whereas in stage I labour is over utilized and capital is underutilized, in stage III the reverse is true; that is, labour is underutilized and capital is over utilized.

In other words, because of the symmetry of production, labour that is over utilized implies that capital is underutilized, and *vice versa*. Since stages I and III of production for labour have been ruled out as illogical from a profit maximization perspective, it also follows that stages III and I of production for capital have been ruled out for the same reasons.

We may infer that stage II of production for labour, and also for capital, is the only region in which production will take place. The precise level of labour and capital usage in stage II in which production will occur cannot be ascertained at this time. For a profit-maximizing firm, the efficient capital–labour combination will depend on the prevailing rental prices of labour (*PL*) and capital (*PK*), and the selling price of a unit of the resulting output (*P*). More precisely, as we will see, the optimal level of labour and capital usage subject to the firm 's operating budget will depend on resource and output prices, and the marginal productivity of productive resources.

**Long run production function**

It was noted earlier that the long run in production describes the situation in which all factors of production are variable. A firm that increases its employment of all factors of production may be said to have increased its scale of operations. Returns to scale refer to the proportional increase in output given some equal proportional increase in all productive inputs. Constant returns to scale (*CRTS*) refer to the condition where output increases in the same proportion as the equal proportional increase in all inputs. Increasing returns to scale (*IRTS*) occur when the increase in output is more than proportional to the equal proportional increase in all inputs. Decreasing returns to scale (*DRTS*) occur when the proportional increase in output is less than proportional increase in all inputs. To illustrate these relationships mathematically, consider the production function

$$Q = f(x_1, x_2, \ldots, x_n)$$ .....*Equation 3.12*

where output is assumed to be a function of *n* productive inputs. A function is said to be homogeneous of degree *r* if, and only if,

$$f(tx_1, tx_2, \ldots, tx_n) \equiv t^r f(x_1, x_2, \ldots, x_n) \quad \text{or} \quad t^r Q \equiv f(tx_1, tx_2, \ldots, tx_n)$$ .....*Equation 3.13*

Where $t > 0$ is some factor of proportionality. Note the identity sign in Equation (3.13). This is not an equation that holds for only a few points but for all $t, x_1, x_2, \ldots, x_n$. This relationship expresses the notion that if all productive inputs are increased by some factor *t*, then output will increase by some factor $t^r$, where $r > 0$. Equation (3.13) is said to be a function that is homogeneous of degree *r*.

Table 3.1. Production functions homogenous of degree r

| R | Returns to scale |
|---|---|
| =1 | Constant |
| >1 | Increasing |
| <1 | Decreasing |

Returns to scale are described as constant, increasing, or decreasing depending on whether the value of *r* is greater than, less than, or equal to unity. Table 3.1 summarizes these relationships. Constant returns to scale is the special case of a production function that is homogeneous of degree one, which is often referred to as *linear homogeneity*.

## 3.2. Theory of Cost

The production function defines the maximum rate of output per unit of time obtainable from a given set of productive inputs. The production function, however, was presented as a purely technological relationship devoid of any behavioural assertions underlying motives of management. The optimal combination of inputs used in the production process will differ depending on the firm ‘s organizational objectives. The objective of profit maximization, for example, may require that the firm use one set of productive resources, while maximization of revenue or market share may require a completely different set. The substitutability of inputs in the production process indicates that any given level of output may be produced with multiple factor combinations. Deciding which of these combinations is optimal not only depends on a well-defined organizational objective but also requires that management attempt to achieve this objective while constrained by a limited operating budget and constellation of factor prices. Changes in the budget constraints or factor prices will alter the optimal combination of inputs. The purpose of this chapter is to bridge the gap between production as a purely technological relationship and the cost of producing a level of output to achieve a well-defined organizational objective.

### 3.2.1. The relationship between production and cost

The cost function of a profit-maximizing firm shows the minimum cost of producing various output levels given market-determined factor prices and the firm ‘s budget constraint. Although largely the domain of accountants, the concept of cost to an economist carries a somewhat different connotation. As already discussed in Chapter 1, economists generally are concerned with any and all costs that are relevant to the production process.

These costs are referred to as *total economic costs*. *Relevant costs* are all costs that pertain to the decision by management to produce a particular good or service.

Total economic costs include the *explicit costs* associated with the day-to- day operations of a firm, but also *implicit* (*indirect*) costs. All costs, both explicit and implicit, are opportunity costs. They are the value of the next best alternative use of a resource. What distinguishes explicit costs from implicit costs is their ―visibility‖ to the manager. Explicit costs are sometimes referred to as ―out-of-pocket‖ costs. Explicit costs are visible expenditures associated with the procurement of the services of a factor of production. Wages paid to workers are an example of an explicit cost.

By contrast, implicit costs are, in a sense, invisible: the manager will not receive an invoice for resources supplied or for services rendered. To understand the distinction, consider the situation of a programmer who is weighing the potential monetary gains from leaving a job at a computer software company to start a consulting business. The programmer must consider not only the potential revenues and out-of-pocket expenses (explicit costs) but also the salary forgone by leaving the computer company. The programmer will receive no bill for the services he or she brings to the consulting company, but the forgone salary is just as real a cost of running a consulting business as the rent paid for office space. As with any opportunity cost, implicit costs represent the value of the factor ‘s next best alternative use and must therefore be considered. As a practical matter, implicit costs are easily made explicit. In the scenario just outlines, the programmer can make the forgone salary explicit by putting himself or herself ―on the books‖ as a salaried employee of the consulting firm.

 **Short-run cost**

The theory of cost is closely related to the underlying production technology. We will begin by assuming that the firm ‘s short-run total cost (*TC*) of production is given by the expression

TC = f(Q).....*Equation 3.14*

The short run in production is defined as that period of time during which at least one factor of production is held at some fixed level. Assuming only two factors of production, capital (*K*) and labour (*L*), and assuming that capital is the fixed factor (*K*0), then Equation (3.14) may be written

$$TC = f[g(K_0, L)]$$ .....*Equation 3.15*

Equation (3.15) simply says that the short-run total cost of production is a function of output, which is itself a function of the level of capital and labour usage. In other words, total cost is a function of output, which is a function of the production technology and factors of production, and factors of production cost money. Equation (3.15) is a general statement that relates the total cost of production to the usage of the factors of production, fixed capital and variable labour. Equation (3.15) also makes clear that total cost is intimately related to the characteristics of the underlying production technology. As we will see, concepts such as total cost (*TC*), average total cost (*ATC*), average variable cost (*AVC*), and marginal cost (*MC*) are defined by their production counterparts, total physical product, average physical product, and marginal physical product of both labour and capital.

To begin with, let us assume that the prices of labour and capital are determined in perfectly competitive factor markets. The short-run total economic cost of production is given as

$$TC = P_K K_0 + P_L L$$ .....*Equation 3.16*

Where $P_K$ is the rental price of capital, $P_L$ is the rental price of labour, $K_0$ is a fixed amount of capital, and $L$ is variable labour input. The most common example of the rental price of labour is the wage rate. An example of the rental price of capital might be what a construction company must pay to lease heavy equipment, such as a bulldozer or a backhoe. If the construction company already owns the heavy equipment, the rental price of capital may be viewed as the forgone income that could have been earned by leasing its own equipment to someone else. In either case, both $P_K$ and $P_L$ are assumed to be market determined and are thus parametric to the output decisions of the firm 's management. Thus, Equation (3.16) may be written

$$TC(Q) = TFC + TVC(Q)$$ .....*Equation 3.17*

Where *TFC* and *TVC* represent total fixed cost and total variable cost, respectively. *Total fixed cost* is a short-run production concept. Fixed costs of production are associated with acquiring and maintaining fixed factors of production. Fixed costs are incurred by the firm regardless of the level of production. Fixed costs are incurred by the firm even if no production takes place at all. Examples often include continuing expenses incurred under a binding contract, such as rental payments on office space, certain insurance payments, and some legal retainers. *Total variable costs* of production are associated with acquiring and maintaining variable factors of production. In stages I and II of production, total variable

cost is an increasing function of the level of output. *Total cost* is the sum of total fixed and total variable cost.

### 3.2.2. Key relationships: average total cost, average fixed cost, average variable cost, and marginal cost

The definitions of average fixed cost, average variable cost, average total cost, and marginal cost are, appropriately, as follows:

$$\text{Average fixed cost: } AFC = \frac{TFC}{Q} \qquad \text{..... } \textit{Equation 3.18}$$

$$\text{Average variable cost: } AVC = \frac{TVC}{Q} \qquad \text{.....} \textit{Equation 3.19}$$

$$\text{Average total cost: } ATC = \frac{TC}{Q} = \frac{TFC + TVC}{Q} = AFC + AVC \qquad \text{.....} \textit{Equation 3.20}$$

$$\text{Marginal cost: } MC = \frac{dTC}{dQ} = \frac{dTVC}{dQ} \qquad \text{.....} \textit{Equation 3.21}$$

*Average total cost* is the total cost of production per unit. It is the total cost of production divided by total output. Average total cost is a short-run production concept if total cost includes fixed cost. It is a long-run production concept if all costs are variable costs. *Average fixed cost*, which is a short-run production concept, is total fixed cost per unit of output. It is total fixed cost divided by total output. *Average variable cost* is total variable cost of production per unit of output. Average variable cost is total variable cost divided by total output. *Marginal cost* is the change in the total cost associated with a change in total output. Contrary to conventional belief, this is not the same thing as the cost of producing the ―last‖ unit of output. Since it is total cost that is changing, the cost of producing the last unit of output is the same as the per-unit cost of producing any other level of output. More specifically, the marginal cost of production for a profit-maximizing firm is equal to average total cost plus the per-unit change in total cost, multiplied by total output. Equation (3.21) shows that marginal cost is the same as marginal variable cost, since total fixed cost is a constant. Related to marginal cost is the more general concept of *incremental cost*. While marginal cost is the change in total cost given a change in output, incremental cost is the change in the firm ‘s total costs that result from the implementation of decisions made by management, such as the introduction of a new product or a change in the firm ‘s advertising campaign. By contrast, *sunk costs* are invariant with respect to changes in management

decisions. Since sunk costs are not recoverable once incurred, they should not be considered when one is determining, say, an optimal level of output or product mix. Suppose, for example, that a textile manufacture purchases a loom for $1 million. If the firm is able to dispose of the loom in the resale market for only $750,000, the firm, in effect, has permanently lost $250,000. In other words, the firm has incurred a sunk cost of $250,000.

Related to the concept of sunk cost is the analytically more important concept of total fixed cost. Total fixed cost, which represents the cost of a firm ‘s fixed inputs, is invariant with respect to the profit-maximizing level of output. This is demonstrated in Equation (3.21). Changes in total cost with respect to changes in output are the same as changes in total variable cost with respect to changes in output. In other words, marginal variable cost is identical to marginal cost. The distinction between sunk and fixed cost is subtle. Suppose that when the firm operated the loom to produce cotton fabrics, the rental price of the loom was $100,000 per year. This rental price is invariant to the firm ‘s level of production. In other words, the firm would rent the loom for $100,000 per year regardless of whether it produced 5,000 or 100,000 yards of cloth during that period. A sunk cost is essentially the difference between the purchase price of the loom and its salvage value.

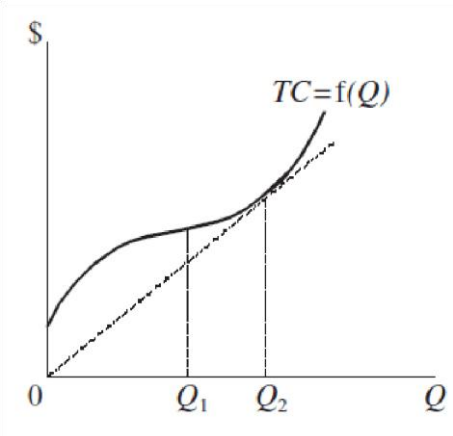


*Figure 3.2 Total cost curve exhibiting increasing then diminishing marginal product*

Additional insights into the relationship between total cost and production may be seen by examining Equation (3.21). Recalling that $Q = f(K, L)$ and applying the chain rule we obtain

$$MC = \frac{dTC}{dQ} = \left(\frac{dTC}{dL}\right)\left(\frac{dL}{dQ}\right)$$ .....*Equation 3.22*

Recalling from Equation (3.16) that $TC = P_K K_0 + P_L L$, then marginal cost may be written as

$$MC = P_L\left(\frac{dL}{dQ}\right) = \frac{P_L}{MP_L}$$ .....*Equation 3.23*

Where $P_L$ is the rental price of homogeneous labour input and *MPL* is the marginal product of labour. Equation (3.23) establishes a direct link between marginal cost and the marginal product of labour, which was discussed in production theory. Since $P_L$ is a constant, it is easily seen that *MC* varies inversely with *MPL*. Recalling Figure 3.1, the shapes of the *TC*, *ATC*, *AVC*, and *MC* curves may easily be explained. When *MPL* is rising (falling), for example, *MC* will be falling (rising). These relationships are illustrated in Figures 3.2 and 3.3. Equation (3.23) indicates clearly the relation between the theory of production and the theory of cost. The cost curves are shaped as they are because the production function exhibits the properties it does, especially the law of diminishing marginal product. In other words, underlying the short-run cost functions are the short-run production functions.

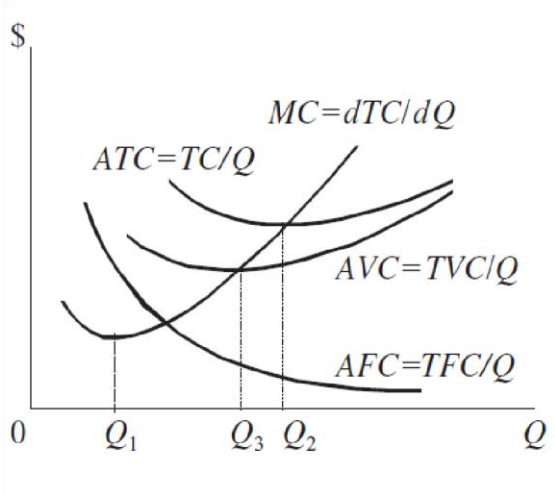


*Figure 3.3 Marginal, average total, average variable and average fixed cost curves*

### 3.2.3. Mathematical relationship between ATC and MC

An examination of Figure 3.3 indicates that the marginal cost (*MC*) curve intersects the average total cost (*ATC*) curve and average variable cost (*AVC*) curve from below at their minimum points. This section will extend that discussion to the specific relation between *MC* and *ATC*. The approach is to minimize the average total cost function and determine if the condition observed in Figure 3.3 is satisfied. Consider, again, the definition of average total cost

$$ATC = \frac{TC}{Q}$$ .....*Equation3.24*

Minimizing Equation (3.24) by taking the first derivative with respect to output and setting the results equal to zero yields

$$\frac{dATC}{dQ} = \frac{Q(dTC/dQ) - TC}{Q^2} = 0$$ .....*Equation 3.25*

Since $Q^2 > 0$, then, from Equation (3.25),

$$\frac{dTC}{dQ} = \frac{TC}{Q}$$ .....*Equation 3.26*

That is, the first-order condition for a minimum *ATC* is

*MC = ATC* .....*Equation 3.27*

The second-order condition for minimum is that the second derivative of Equation (3.27) is positive. Taking the derivative of Equation (3.25) yields

$$\frac{d^2ATC}{dQ^2}$$

$$= \frac{Q^2[Q(d^2TC/dQ^2 + dTC/dQ - dTC/dQ)] - [Q(dTC/dQ) - TC](2Q)}{Q^4}$$

$$= \frac{(d^2TC/dQ^2)}{Q} - \frac{(2dTC/dQ)}{Q^2} + \frac{2TC}{Q^3}$$

Substituting the first-order condition $Q(dTC/dQ) - TC$ yields

$$\frac{d^2ATC}{dQ^2} = \frac{(d^2TC/dQ^2)}{Q} - \frac{(2dTC/dQ)}{Q^2} + \frac{2Q(dTC/dQ)}{Q^3}$$

$$= \frac{(d^2TC/dQ^2)}{Q} = \frac{(dMC/dQ)}{Q} > 0$$ .....*Equation 3.28*

for a minimum *ATC*. Since $Q > 0$, at a minimum *ATC* the second-order condition requires that marginal cost is increasing. As indicated in Figure 3,3 at the point at which *ATC* is minimized, *ATC = MC*, and the *MC* curve intersects the *ATC* curve from below.

It is also easy to demonstrate that the marginal cost curve also intersects the minimum point on the average variable cost (*AVC*) curve from below. From Equation (3.17)

*TC*(*Q*) =*TFC* +*TVC*(*Q*)

Dividing both sides of Equation (2.17) by output, and rearranging, we obtain

*AVC*(*Q*) = *ATC*(*Q*) – *AFC*.....*Equation 3.29*

Taking the derivative of Equation (3.29) with respect to output yields

$$\frac{dAVC(Q)}{dQ} = \frac{dATC(Q)}{dQ}$$ .....*Equation 3.30*

since *AFC* is a constant. Thus, minimizing average variable cost with respect to output is equivalent to minimizing average total cost with respect to output, and generates identical conclusions. It should be pointed out that although the *MC* curve intersects both the *ATC* and *AVC* curve at their minimum points that minimum *AVC* occurs at a lower output level than minimum *ATC*. To see why this is the case, refer again to Equation (3.29). For any output level, including the output level that minimizes *ATC* and *AVC*, *ATC* exceeds *AVC* by the amount of *AFC*. In Figure 3.3, since the marginal cost curve is upward sloping, the *MC* curve cannot intersect both the *AVC* and *ATC* curves at their minimum points unless it does so at different output levels as long as $AFC > 0$ (otherwise $ATC = AVC$). Since $ATC > AVC$, the output level corresponding to the minimum point on the *ATC* curve must be above and to the right of the output level that corresponds to the minimum point on the *AVC* curve. Finally, referring again to Figure 3.2, from Equation (3.29), *AFC* must equal the vertical distance between the *ATC* and the *AVC* curves at any output level. Since *AFC* falls as *Q* increases, the *ATC* and *AVC* curves must be asymptotic vertically.

### 3.2.4. Learning curve effect

The discussion about production noted that the profit-maximizing firm will operate in *stage II* of production. It will be recalled that in *stage II* the marginal product of a factor, say labor, is positive, but declining at an increasing rate. The phenomenon is a direct consequence of the law of diminishing marginal product. At constant factor prices this relationship implies that as output is expanded, the marginal cost of a variable factor increases at an increasing rate. An important assumption implicit in the law of diminishing marginal product is that the quality of the variable input used remains unchanged.

***Table 3.2. Learning Curve Effects: Unit Labour Costs***

| Output | $m = 0.9$ | $m = 0.8$ | $m = 0.7$ |
|---|---|---|---|
| 1 | $10,000.00 | $10,000.00 | $10,000.00 |
| 2 | 9,000.00 | 8,000.00 | 7,000.00 |
| 4 | 8,100.00 | 6,400.00 | 4,900.00 |
| 8 | 7,290.00 | 5,120.00 | 3,430.00 |
| 16 | 6,561.00 | 4,096.00 | 2,401.00 |
| 32 | 5,904.90 | 3,276.80 | 1,680.70 |
| 64 | 5,314.41 | 2,621.44 | 1,176.49 |
| 128 | 4,782.97 | 2,097.15 | 823.54 |
| 256 | 4,340.67 | 1,677.72 | 576.48 |

In the case of labour, for example, the ―productivity‖ of labour is assumed to remain unchanged regardless of the level of production. In fact, this restriction is an

oversimplification: it is reasonable to expect that as output expands over time, the typical labourer ―gets better‖ at his or her job. In other words, it is reasonable to assume that workers become more productive as they gain experience. This would suggest that over some range of production, per-unit labour input might, in fact, fall. At constant labour prices, this implies that per-unit labour costs may in fact fall. We are not, of course, talking about *stage I* of production, where per-unit costs fall because of increased specialization as additional units of, say, labour are added to underutilized amounts of capital. On-the-job training and experience will make workers more productive, which has important implications for the cost structure of the firm. It is precisely the expectation of greater productivity that compels many firms to underwrite on-site employee training and off-site continuing education programs. The relationship between increased per worker productivity and reduced per worker cost at fixed labour prices associated with an increase in output and experience is called the *learning curve effect*.

*Definition:* The learning curve effect measures the relation between an increase in per-worker productivity (a decrease in per-unit labour cost at fixed labour prices) associated with an improvement in labour skills from on the- job experience. One measure of the learning curve effect is

$$\Gamma = \text{Per-unit labor cost} = \varphi Q^{\beta}$$ .....*Equation 3.31*

where $\varphi$ is the per-unit cost of producing the first unit of output, $Q$ is the level of output, $\beta = (\ln m)/(\ln \lambda)$, $m$ is the learning factor, and $\lambda$ is a factor of output proportionality. In fact, b is a measure of the learning curve effect. It determines the rate at which per-unit labour requirements fall given the rate at which workers ―learn‖ ($m$) following a scalar (proportional) increase in production ($\lambda$). The value of $m$ varies from zero to unity ($0 \leq m \leq 1$). As the value of $m$ approaches zero, the learning curve effect on lowering per unit labour costs becomes more powerful. Table 3.2 presents data for 90% ($m = 0.9$), 80% ($m = 0.8$), and 70% ($m = 0.7$) learning curves for a factor of output proportionality ($\lambda$) of 2. That is, each time output is doubled; the per unit labour cost drops to become 90, 80, or 70% of its previous level. It is assumed that it takes 1,000 labour units (say, labour hours) to produce the first unit of output, and that the wage rate is constant at $10 per hour. Thus, the per-unit labour cost of producing the first unit of output is $10,000. If the learning factor is $m = 0.9$ (i.e., the learning process is relatively slow), the per-unit labour cost when $Q = 2$ is $9,000. When $Q = 4$, per unit labour cost is $8,100, and so on. The reader will note that

the lower the learning factor (i.e., the quicker the learning process), the more rapidly per-unit labour costs fall as output expands. Since the wage rate in this example is constant at $10 per hour, the learning curve effect implies that fewer labour units per unit of output are required as output increases. The extreme cases are $m = 1$ and $m = 0$. When $m = 1$, no learning whatsoever takes place, and $\beta = (\ln 1 / \ln \lambda) = 0$. In this case, there are no learning curve effects, and the per-unit labour cost remains unchanged. On the other hand, when $m = 0$, then $\beta = -\infty$. In this case, learning is so complete and production so efficient that the per-unit labour cost reduces to zero, which implies that at a constant wage rate no labour, is required at all. In the two-input case, this suggests that all production technology is embodied in the amount of capital employed.

Learning curve effects are usually thought to result from the development of labour skills, especially for tasks of a repetitive nature, over time. More broadly, however, incremental reductions in per-unit labour costs may result from a variety of factors, such as the adoption of new production, organizational, and managerial techniques, the replacement of higher cost with lower cost materials, an new product design. Consideration of these additional factors has given rise to the broader term *experience curve effects*.

*Definition:* The experience curve effect is a measure of the relationship between an increase in per-worker productivity (a decrease in per-worker cost at fixed labour prices) associated with an improvement in labour skills from on-the-job experience, the adoption of new production, organizational and managerial techniques, the replacement of higher cost with lower cost materials, new product design, and so on.

- **Long-run cost**

In the long run all factors of production are assumed to be variable. Since there are no fixed inputs, there are no fixed costs. All costs are variable. Unlike the short-run production function, however, there is little that can be said about production in the long run. There is no long-run equivalent of the law of diminishing marginal product. As in the case of the firm ‘s short-run cost functions, long-run cost functions are intimately related to the long-run production function. In particular, the firm ‘s long-run cost functions are related to the concept of returns to scale, which was discussed in production theory. In general, economists have theorized that an increase in the firm ‘s scale of operations (i.e., a proportional increase in all inputs), is likely to be accompanied by increasing, constant, and decreasing returns to scale. The relation between total output and all inputs, the *long-run total product curve* (*LRTP*) is illustrated in Figure 3.4.

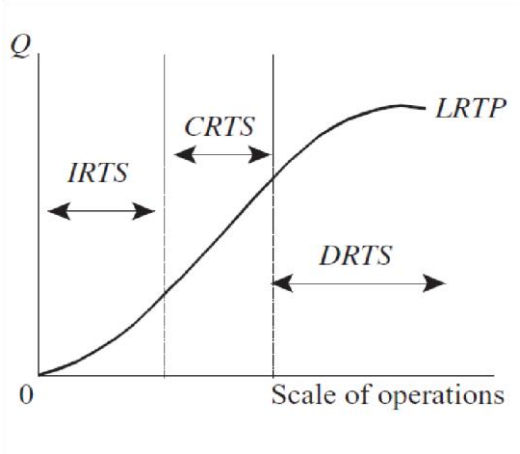


*Figure 3.4. Long-run total product curve*

It can be recalled from our discussion on production that the coefficient of output elasticity for increasing returns to scale, constant returns to scale, and decreasing returns to scale, which is the sum of the output elasticities of each input, is greater than unity, equal to unity, and less than unity, respectively. Although the shapes of the long-run and short-run total product curves are similar, the reasons are quite different. The short-run total product curve derives its shape from the law of diminishing marginal product.

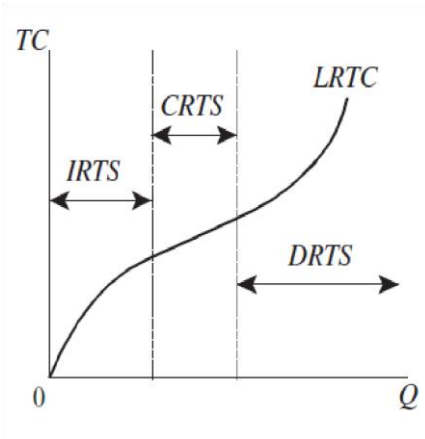


*Figure 3.5. Long run total cost curve*

The long-run total product curve derives its shape from quite different considerations. To begin with, firms might experience increasing returns to scale during the early stages of production because of the opportunity for increased specialization of both human and nonhuman factors, which would lead to gains in efficiency. Only large firms, for example, can rationalize the use of in-house lawyers, accountants, economists, and so on. Another reason is that equipment of larger capacity may be more efficient than machinery of smaller capacity.

Eventually, as the firm gets larger, the efficiencies from increased size may be exhausted. As a firm grows larger, so too do the demands on management. Increased administrative layers may become necessary, resulting in a loss of efficiency as internal coordination of production processes become more difficult. At some large level of output, factors of production may become overworked and diminishing returns to management may set in. Problems of interdepartmental and interdivisional coordination may become endemic. The result would be decreasing returns to scale. Since there are no fixed costs in the long run, the firm ‘s corresponding *long-run total cost curve* (*LRTC*), which is similar in appearance to the short run total cost curve, intersects *TC* at the origin. This is illustrated in Figure 3.5.

### 3.2.6. Economies of scale

*Scale* refers to size. What is the effect on the firm ‘s per-unit cost of production following an increase in all factors of production? In other words, does the size of a firm ‘s operations affect its per-unit cost of production?

The answer to the second questions is that it may. Regardless of what we have learned about the short-run production function and the law of diminishing marginal product, it is difficult to generalize about the effect of size on per-unit cost of production, although we may speculate about the most likely possibilities.

*Economies of scale* are intimately related to the concept of increasing returns to scale. If per unit costs of production decline as the scale of a firm ‘s operations increase, the firm is said to experience economies of scale. The reason is simple. If we assume constant factor prices, while the firm ‘s total cost of production rises proportionately with an increase in total factor usage, per-unit cost of production falls because output has increased more than proportionately. In other words, an increase in the firm ‘s scale of operations results in a decline in *long-run average total cost* (*LRATC*).

Conversely, *diseconomies of scale* are intimately related to the concept of decreasing returns to scale. Once again, the explanation is straightforward. Assuming constant factor prices, the firm ‘s total cost of production rises proportionately with an increase in total factor usage, but the per-unit cost of production increases because output increases less than proportionately.

In other words, an increase in the firm ‘s scale of operations results in an increase in the firm ‘s long-run average total cost.

Finally, in the case of *constant returns to scale*, per-unit cost of production remains constant as production increases or decreases proportionately with an increase or decrease in factor usage. In other words, an increase or decrease in the firm ‘s size will have no effect on the firm ‘s long-run average total cost.

*Definition:* Assuming constant factor prices, economies of scale occur when per-unit costs of production decline following a proportional increase in all factors of production.

Definition: Assuming constant factor prices, diseconomies of scale occur when per-unit costs of production increase following a proportional increase in all factors of production.

Assume that $Q = f(L, K)$, where $Q$ is output, $L$ is variable labour, and $K$ is fixed capital. Figure 3.6 illustrates several possible *short-run average total cost curves* (*SRATC*), each corresponding to a different level of capital usage. Here, $SRATC_2$ represents the short-run average total cost curve of the firm utilizing a higher level of fixed capital input than $SRATC_1$, $SRATC_3$ represents the short-run average total cost curve of the firm utilizing a higher level of fixed capital input than $SRATC_2$, and so on. We can see that the firm may produce a given level of output with one or more short-run production functions: for example, the firm may produce output level $Q_1$ along $SRATC_1$ (point *A*) or $SRATC_2$ (point *B*). If the firm believes that the demand for its product in the foreseeable future is $Q_1$, it will choose to employ a level of capital consistent with $SRATC_2$ to realize lower per-unit cost of production. As the demand for the firm ‘s output increases, the firm will choose that production technology such that its per-unit costs are minimized. All such points are illustrated by the ―envelope‖ of the short-run average total cost curves. This curve is called the *long-run average total cost curve*, which is labelled *LRATC* in Figure 3.6.

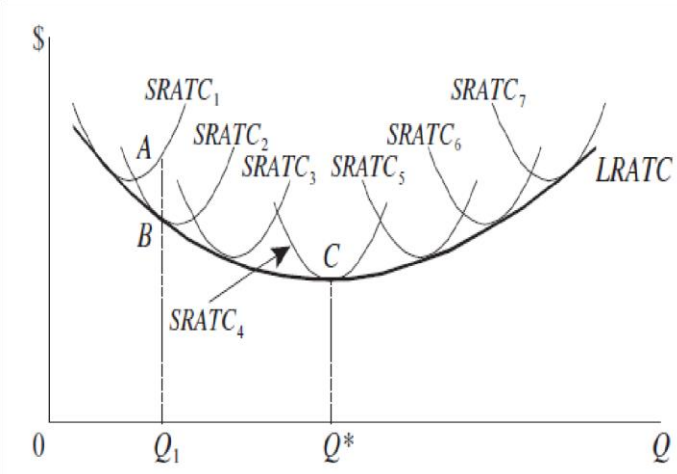


*Figure 3.6. Long-run average total cost curve as the “envelope” of the short-run average total cost curves.*

Unlike the short-run average total cost curves, which derive their shape from the law of diminishing returns, the long-run average total cost curve derives its shape from the returns to scale characteristics of the underlying production function. In Figure 6.6, the range of output levels 0 to $Q^*$ corresponds to the underlying production characteristic of increasing returns to scale. Over this range of output, *LRATC* is downward sloping (i.e., $dLRATC/dQ < 0$). Output levels greater than $Q^*$ reflect decreasing returns to scale. Over this range of output *LRATC* is upward sloping (i.e., $dLRATC/dQ > 0$). Where the long-run average total cost curve is neither rising nor falling, $dLRATC/dQ = 0$, the underlying production function exhibits constant returns to scale.

The output level that corresponds to minimum long-run ATC is commonly referred to as the *minimum efficient scale* (*MES*). Minimum efficient scale is the level of output that corresponds to the lowest per-unit cost of production in the long run.

The shape of the long-run average total cost curve will vary from industry to industry. There is, in fact, no *a priori reason* for the *LRATC* curve to be U shaped. Figures 3.7 through 3.9 illustrate three other possible shapes of the long-run average total cost curve.

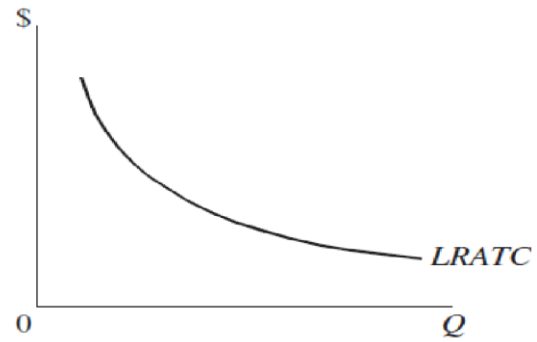


Figure 3.7. Economies of Scale

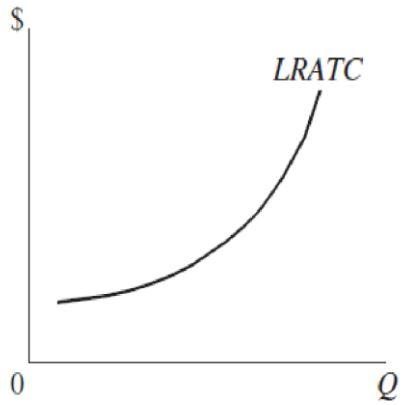


*Figure 3.9. Diseconomies of Scale*

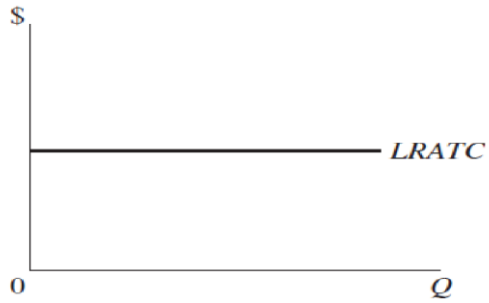


Figure 3.8. Constant Returns to Scale

 Reasons for economies and diseconomies of scale

As firms grow larger, economies of scale may be realized as a result of specialization in production, sales, marketing, research and development, and other areas. Specialization may increase the firm ‘s productivity in greater proportion than the increase in operating cost associated with greater size. Some types of machinery, for example, are more efficient for large production runs. Examples include large electrical-power generators and large blast furnaces that generate greater output than smaller ones. Large companies may be able to extract more favourable financing terms from creditors than small companies.

Large size, however, does not guarantee that improvement in efficiency and lower per-unit costs. Large size is often accompanied by a disproportionately great increase in specialized management. Coordination and communication between and among departments becomes more complicated, difficult, and time-consuming. As a result, time that would be better spent in the actual process of production declines and overhead expenses increase disproportionately. The result is that diminishing returns to scale set in, and per-unit costs rise. In other words, growth usually is accompanied by diseconomies of scale.

### 3.2.7. Multiproduct cost functions

We have thus far discussed manufacturing processes involving the production of a single output. Yet, many firms use the same production facilities to produce multiple products. Automobile companies, such as Ford Motor Company, produce both cars and trucks at the same production facilities. Chemical and pharmaceutical companies, such as Dow Chemical and Merck, use the same basic production facilities to produce multiple different products. Computer companies, such as IBM, produce monitors, printers, scanners, modems and, of course, computers. Consumer product companies, such as General Electric, produce a wide range of household durables, such refrigerators, ovens, and light-bulbs. In each of these examples it is reasonable to suppose that total cost of production is a function of more than a single output. In the case of a firm producing two products, the *multiproduct cost function* may be written as

$$TC = TC(Q_1, Q_2)$$ .....*Equation 3.32*

Where $Q_1$ and $Q_2$ represent the number of units produced of goods 1 and 2, respectively. Definition: A multiproduct cost function summarizes the cost of efficiently using the same production facilities to produce two or more products.

The multiproduct cost function summarized in Equation (3.32) has the same basic interpretation as a single product cost function, although the specific relationship depends on the manner in which these goods are produced. Two examples of multiproduct cost relationships are *economies of scope* and *cost complementarities*.

### 3.2.8. Economies of scope

Economies of scope exist when the total cost of using the same production facilities to produce two or more goods is less than that of producing these goods at separate production facilities. In the case of a firm that produces two goods, economies of scope exist when

$$TC(Q_1,0)+TC(0,Q_2)>TC(Q_1,Q_2)$$ .....*Equation 3.33*

It may be less expensive, for example, for Ford Motor Company to produce cars and trucks by more intensively utilizing a single assembly plant than to produce these products at two separates, less intensively utilized, manufacturing facilities. It is less expensive for the restaurants in the Olive Garden chain to use the same ovens, tables, refrigerators, and so on to produce both pasta and parmigiana meals than to duplicate the factors of production.

*Definition:* Economies of scope in production exist when the total cost of producing two or more goods is lower when the same production facilities are used than when separate production facilities are used to produce the goods.

### 3.2.9. Cost complementarities

Cost complementarities exist when the marginal cost of producing $Q_1$ is reduced by increasing the production of $Q_2$. From Equation (3.32), the marginal cost of producing $Q_1$ may be written as

$$MC_1(Q_1,Q_2)=\frac{\partial TC}{\partial Q_1}$$

The multiproduct cost function exhibits cost complementarities if the cross-second partial derivative of the multiproduct cost function is negative, that is,

$$\frac{\partial MC_1(Q_1,Q_2)}{\partial Q_2}=\frac{\partial^2 TC}{\partial Q_1 \partial Q_2}<0$$ .....*Equation 3.34 a*

$$\frac{\partial MC_2(Q_1,Q_2)}{\partial Q_1}=\frac{\partial^2 TC}{\partial Q_2 \partial Q_1}<0$$ .....*Equation 3.34 b*

Equations (3.34) says that an increase in the production of $Q_2$ reduces the marginal cost of producing $Q_1$.

*Definition:* Cost complementarities exist when the marginal cost of producing one good is reduced by increasing the production of another good.

Examples of cost complementarities include the simultaneous production of beef and leather, and doughnuts and doughnut ―holes. Slaughtering cattle for beef without considering the potential usefulness of leather by-products, and vice versa, is not only poor business but wasteful. Producing doughnuts and doughnut holes is certainly more cost efficient when these snacks are produced together than when they are produced separately.

***Self Assessment exercise***

a.A firm can manufacture a product according to the production function $Q = F(K, L) = K^{3/4}L^{1/4}$ Calculate the average product of labor, $AP_L$, when the level of capital is fixed at 16 units and the firm uses 16 units of labor. How does the average product of labor change when the firm uses 81 units of labor?

b. Find an expression for the marginal product of labor, $MP_L$, when the amount of capital is fixed at 16 units. Then, illustrate that the marginal product of labor depends on the amount of labor hired by calculating the marginal product of labor for 16 and 81 units of labor.

c.Suppose capital is fixed at 16 units. If the firm can sell its output at a price of $100 per unit and can hire labor at $25 per unit, how many units of labor should the firm hire in order to maximize profits?

1. Explain the difference between the law of diminishing marginal returns and the law of diminishing marginal rate of technical substitution.

2. An economist estimated that the cost function of a single-product firm is

$$C(Q) = 50 + 25Q + 30Q^2 + 5Q^3$$

Based on this information, determine:

a. The fixed cost of producing 10 units of output.
b. The variable cost of producing 10 units of output.
c. The total cost of producing 10 units of output.
d. The average fixed cost of producing 10 units of output.
e. The average variable cost of producing 10 units of output.
f. The average total cost of producing 10 units of output.
g. The marginal cost when $Q = 10$.

# CHAPTER FOUR
# 4. PRICING DECISIONS

***Objectives***

- *To explain the concept of demand and demand analysis*
- *To understand how to apply simple elasticity-based mark up formulas to determine profit-maximizing prices in environments where a business enjoys market power*
- *To explain how to formulate pricing strategies that permit firms to extract additional surplus from consumers*

## 4.1. Demand analysis

It is of vital importance for any firm to understand the demand for its products. Demand relationships determine revenues and indirectly affect output and costs; they thus have a fundamental impact on profits. An understanding of demand is also relevant for planning purposes, involving production, transportation, inventory, sales, marketing and finance functions. The identification of factors affecting demand and the precise effects that these have is therefore a core element in managerial economics.

Unfortunately, the word _demand ‘can be used in a variety of senses, which often causes misunderstanding and errors of analysis. We can talk about demand curves, schedules, functions, equations, relationships or points. When economists use the single word _demand ‘they are referring to the relationship that is frequently called the demand curve. In this sense, demand refers to the quantities that people are or would be willing to buy at different prices during a given time period, assuming that other factors affecting these quantities remain the same. It is worth noting that this definition incorporates three important concepts:

i. It involves three parameters – price, quantity and time.
ii. It refers to quantities in the plural, therefore a whole relationship, not a single quantity.
iii. It involves the ceteris paribus (other things being equal) assumption, which is a very common one in making statements in economics.

The second of the above concepts, in particular, tends to cause misunderstandings, since it is common in both business and marketing to refer to demand as a single quantity. For example, consider the following statement: the demand for Coke is 80 cans per day at the price of 50 pence per can. First of all, there are issues of definition and measurement. Whose demand is being considered? Does _Coke ‘include Diet Coke and different flavours? What

is the size of the can? How do we include bottles? However, even if these issues are clarified, the statement is still technically incorrect from an economist ‘s standpoint.

The statement refers to a quantity demanded, not to demand. As it will be seen, it is vital to distinguish between these two concepts, as otherwise errors in analysis will occur. It is now useful to consider the different ways of expressing the demand for a product since this is helpful in illustrating this distinction.

- **Tables, graphs and equations**

**a. Tables**

These are the simplest method of representation. An example of a demand table or schedule faced by a retailer is given in Table 4.1. The table shows the general _law of demand ‘, that less is demanded at higher prices. It should be noted that the original statement above relating to a quantity demanded is included in the table as one of the five pairs of values. Although they are simple to understand the main problem with tables is that they are not very useful for analytical purposes. *Table 4.1. Demand table*

| Price of Coke (pence per can) | Quantity sold (cans per day) |
|---|---|
| 30 | 120 |
| 40 | 100 |
| 50 | 80 |
| 60 | 60 |
| 70 | 40 |

**b. Graphs**

These are much more useful for analysis, and indeed most analysis in introductory microeconomics involves the use of graphs. A graph relating to the above demand schedule is shown in Figure 4.1. It can be seen that the demand relationship in this case is both inverse and linear. At this stage both of these characteristics are assumed, but later on both of these assumptions will have to be examined. Again, the difference between the concepts of demand and quantity demanded is illustrated: the former relates to the whole demand curve whereas the latter relates to a single point on the curve.

Although graphical analysis is very useful in economics its main disadvantage is that it essentially involves a two-dimensional framework. Thus, it is mainly limited to examining two-variable relationships. Demand relationships often involve many variables and although the effects of these can be shown on a graph, they are difficult to measure.

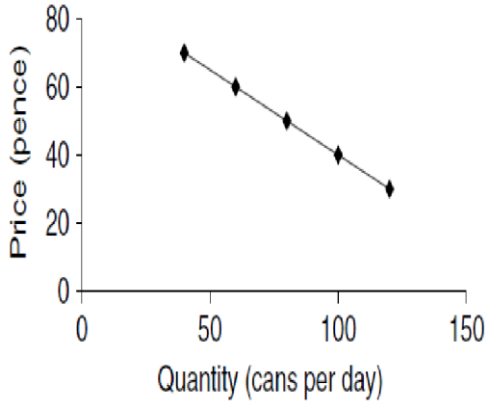


*Figure 4.1. Demand graph*

**c. Equations**

These are the most useful method of representation for analytical purposes since they explicitly show the effects of all the variables affecting quantity demanded, in a concise form that at the same time reveals important information regarding the nature of the relationship. Taking the demand curve in Figure 4.1, we can estimate the equation relating to it in various ways. The general form of the demand function in terms of price and quantity demanded is:

$Q = f(P)$.....*Equation 4.1*

This is the most general way of describing the relationship since it does not involve any specific mathematical form. It can obviously be expanded by including any number of variables that might affect quantity demanded on the right-hand side of the equation, for example:

$Q = f(P, A, Y, P_s)$ .....*Equation 4.2*

where A represents advertising expenditure, Y represents average income of the market and $P_s$ represents the price of a substitute product.

In the two-variable case the demand function can be expressed in a linear form:

$Q = a + bP$.....*Equation 4.3*

The coefficients a and b can then be calculated for the demand schedule in Table 4.1. One way of doing this is to use simultaneous equations and substitute any two pairs of values in the table to solve for a and b. However, it is more insightful to calculate the value of b first, using the mathematical concept that b represents $\Delta Q / \Delta P$. Again, any two pairs of values can be used to establish that b= -2. For example, if the first two pairs of values are taken, b

= -20/10 = -2. Then any pair of values can be used, substituting the value of b, to calculate the value of a.

- **Demand and revenue**

The demand curve for a firm automatically determines its revenue function, since revenue is simply price multiplied by quantity. *Table 4.2. Demand and Revenue relationship*

| Quantity | Price | Total revenue | Marginal revenue |
|---|---|---|---|
| 40 | 70 | 2,800 | 50 |
| | | | **40** |
| 60 | 60 | 3,600 | 30 |
| | | | **20** |
| 80 | 50 | 4,000 | 10 |
| | | | **0** |
| 100 | 40 | 4,000 | –10 |
| | | | **–20** |
| 120 | 30 | 3,600 | –30 |

We can now rearrange and expand the information in Table 4.1 to show how total and marginal revenue can be measured. The table is rearranged because revenue and marginal revenue are normally considered as functions of quantity rather than price. It should be noted that marginal revenue can be calculated in two different ways, which must not be confused:

1. $\Delta R=\Delta Q$ – Thus when quantity changes from 40 to 60

MR = (3600 – 2800)/ (60 – 40) = 40

These values are given in bold in Table 4.2 and inserted between the quantities of 40 and 60.

2. dR/dQ – This is the derivative of the revenue function.

Since P = 90 - 0.5Q

$R = PQ = 90\ Q = 0.5Q^2$

MR = 90 - Q

These values are given by substituting the value of Q into the MR function and are written directly opposite the relevant values of Q in the table.

- **Factors determining demand**

This section is concerned with identifying the factors which affect the quantity demanded of a product, describing the nature of these factors, examining how they can be measured, and examining the relationship with quantity demanded. These factors can be considered in

terms of a demand function: Q = f (P, L, A, D, Y) where each symbol on the right-hand side denotes a relevant factor. P refers to price, L to quality, A to advertising spending, D to distribution spending and Y to the average income of the market.

It is useful from a managerial decision-making viewpoint to distinguish between controllable and uncontrollable factors. Controllable in this context means controllable by the firm; the distinction in practice between what can be controlled and what cannot is somewhat blurred, as will be seen. The relevant factors are summarized in Figure 4.2 below.

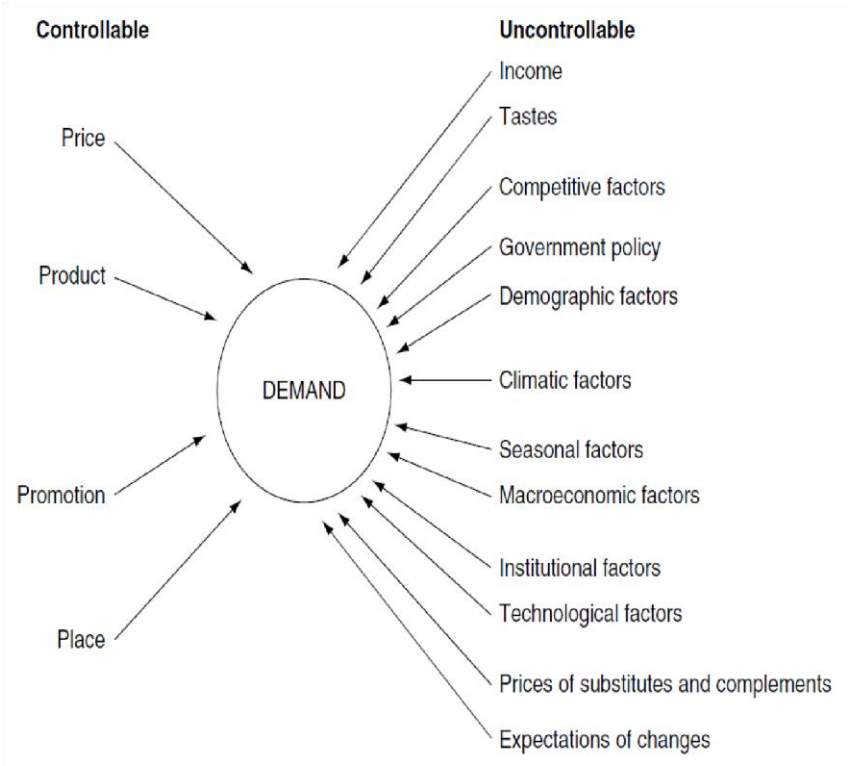


*Figure 4.2. Factors determining demand*

- **Controllable factors**

These correspond to what are often referred to as the marketing-mix variables.

**a. Price**

The nature of price can affect demand positively or negatively depending on different factors. The firm may well be a multiproduct firm, in which case the prices of its other products are also relevant, since they are often substituting or complements. This factor is discussed in the context of _uncontrollable factors '.

**b. Product**

It is the perceived quality of the product that affects quantity demanded. In practice there are two problems of measurement. The first of these is subjectivity: different consumers perceive quality differently. The second is multidimensionality which is now discussed at

greater length. Products often have many different attributes, as was mentioned in the previous section. This is particularly obvious for expensive and hi-tech products. With mobile phones, for example, a consumer may be looking for a good design, small size and weight, multiple features, a good complementary network, a good brand name, a good warranty, and so on. Even with an inexpensive product like toothpaste, consumers may be looking not just for something to clean their teeth, but also something that whitens their teeth, smells fresh, prevents decay and plaque and comes in a convenient dispenser.

Marketing professionals and researchers often make use of multidimensional scaling in this situation. Consumers are asked to rate a product on a range of characteristics, giving a score for each one. Such questionnaires have to be carefully designed in order to include all the relevant factors; they also involve assumptions regarding the relative importance of the factors in terms of their weighting. Sometimes the unit cost of production is used as an approximate measure of quality; this again may not be an accurate indicator of quality, since it may be more a reflection of the efficiency of the firm.

**c. Promotion**

This refers to communication by the firm with its intended customers, ultimately aimed at persuading them to buy the product. The main components of the promotion-mix are customarily identified as: advertising, personal selling, publicity and sales promotion. Sometimes other elements are included separately, like direct marketing or packaging. These are usually measured in terms of expenditure and there is a direct relationship with quantity demanded.

**d. Place**

This is not a very apt description of what is involved here. It refers to distribution strategy, in particular the selection of distribution channels. This is again measured in terms of expenditure and again there is a direct relationship with quantity demanded.

In summary, the last three components of the marketing mix can all be considered as costs, while price is the only component that contributes directly to revenue; its relationship with revenue depends on price elasticity.

- **Uncontrollable factors**

These generally are external or environmental, and include a wide variety of factors whose relative importance varies from one product to another.

**a. Income**

This is a very important factor affecting quantity demanded for the majority of goods. Although it is expressed numerically there are a number of measurement problems. First of all, economists measure income in a number of different ways. The measure that is usually the most relevant to determining demand is personal disposable income (PDI). Although most governments or governmental agencies publish reliable figures relating to this on a regular basis, these are on a national basis and are therefore only strictly applicable to firms that have national markets. It is the average personal disposable income of the market for the firm that is relevant. There is thus a measurement problem if the firm has only a small market, because of the difficulty in obtaining relevant information; this is easier for national markets, or at least regional markets. In the situation where a firm has a small market, or one that is difficult to identify in terms of geographical area or otherwise, some proxy income variable may have to be used. For example, a shop selling products to just a small area or customer base within Addis Ababa may have to use average income in Addis Ababa as a proxy variable. The relationship with quantity demanded can be direct or inverse depending on whether the product is normal or inferior.

**b. Tastes**

These are difficult to define and measure, because they relate to subjective preferences and are multidimensional, just like perceived quality, which they influence. They can vary considerably both inter temporally and inter spatially. Inter temporal variations are variations over time; these apply especially to products whose demand is affected by fashion. This refers not just to clothes, but to other consumer products like entertainment, food and sports. Inter-spatial variations are variations from place to place, especially from country to country. Much of this variation arises because of socio-cultural factors. Of course, tastes are not an entirely uncontrollable factor; firms try to influence tastes through their advertising. Cigarette advertising is an example of this, and this is one factor that causes governments to regulate it. **c. Government policy**

This has both micro- and macroeconomic effects. The first of these are considered here. Governments often want to discourage us from buying certain products and sometimes to encourage us to buy more of other products. Examples of the first category are cigarettes, alcohol, many other drugs (as opposed to medicines, which are in the second category, even though in practice the distinction may be very difficult to draw), and weapons, products whose consumption or production causes environmental damage, and sometimes

pornography and foreign goods. Goods in the second category are often their opposites. There are many policy instruments in either case: bans, indirect taxes/subsidies, health warnings, promotion regulations, licences, age restrictions, restrictions on times and places of consumption. It is important to distinguish between policies primarily affecting supply, like the first two, and those affecting demand.

**d. Competitive factors**

This refers to the marketing mix of competitors. Price is not the only relevant factor here because firms often compete in other ways. This non-price competition occurs particularly in oligopolistic markets. An example of this is the competition between Coke and Pepsi, who have a virtual duopoly of the cola market. The competition between them occurs mainly in terms of promotion and distribution.

**e. Demographic factors**

These refer not only to the size of population, but also to its structure. In particular the ageing of the population affects demand in many ways because older people demand more health services and pensions. This has important implications for government policy and the taxpayer.

**f. Climatic factors**

Weather, rainfall, temperature and also terrain are important. In Ethiopia, for example, there is a greater demand for jackets in highland areas.

**g. Seasonal factors**

These refer to any regular, repeated pattern in demand. Many products have a demand that varies according to season of the year, like air travel, hotels, car rental, jewellery and restaurants. The latter also has a monthly pattern (more people eat out at the beginning of the month when they have more cash), a weekly pattern (more people eat out at weekends) and a daily pattern. Many products have a daily pattern in demand, for example public transport, electricity, telephone calls and health clubs. This also has implications for pricing; many firms use peak and off-peak pricing.

h. Macroeconomic factors

These include income, discussed earlier, and also interest rates, exchange rates, inflation rates and unemployment rates. If interest rates fall, for example, this may affect demand for two reasons: many homeowners have mortgages and the falling interest rate will increase their discretionary income. This is the income that they have available to buy non-necessities. Thus, they will buy more of most (normal) goods. In addition, the demand for

consumer durables will be increased, since these are often bought on credit and will therefore be in effect cheaper. Government policy obviously affects these macroeconomic factors to some degree.

**i. Institutional factors**

These include a wide variety of effects; the physical infrastructure of roads, railways and telecommunications is relevant, so are political systems, legal systems, education systems, housing systems, religious systems and family systems. For example, a country that has a poor transportation infrastructure will have a low demand for cars; a country with low literacy levels will have a relatively low demand for newspapers and magazines.

**j. Technological factors**

These primarily affect supply but also affect demand indirectly; this is best explained with an example. Mobile phones will again serve this purpose. Before the 1980s the concept of a truly mobile phone was still in the realm of science fiction, because the technology was not available. Thus, consumers had not taken time to consider the potential of such a product and there was no real demand. As advancing technology made possible smaller, lighter, more effective and cheaper products, consumers had a true chance to evaluate the potential of the product and then demand began to increase accordingly.

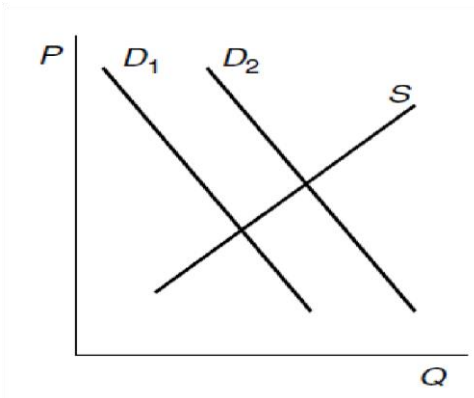


*Figure 4.3. Change in demand*

k. Price of substitutes and complements

A substitute is a product with similar perceived functions or attributes; cars and public transport are substitutes. If the price of public transport were to rise this would increase the demand for cars, other things being equal. Complementary products are consumed together in some way; cars and petrol are complements. If the price of petrol were to rise this would

reduce the demand for cars, other things being equal. Thus, in the case of substitutes the relationship is positive, while in the case of complements the relationship is negative.

### 4.1.1. The Elasticity concepts

Elasticity in general terms is concerned with the responsiveness or sensitivity of one variable to changes in another. Demand elasticity is concerned with how the quantity demanded is responsive to changes in any of the factors of demand. Essentially elasticities are an alternative way of measuring responsiveness to the marginal effects; however, they have the advantage that they allow more meaningful comparisons between different products because they are measured in relative terms, meaning percentages, rather than in absolute units. Theoretically we could discuss elasticities for all the different controllable and uncontrollable factors affecting demand but in practice there are four main types of elasticity that tend to be measured and examined, corresponding to four particularly important factors in the demand function: **price, promotion, income** and **the price of a related product**:

$Q = f(P, A, Y, P')$ .....*Equation 4.4*

Each of the above elasticities needs to be analysed in terms of: definition, measurement, range of values, determinants, use and applications.

*4.1.1.1. Own price elasticity*

- ***Definition***

PED (Price Elasticity of Demand) is the percentage change in quantity demanded in response to a 1 per cent change in price. In symbols we can write:

$PED = \%\Delta Q / \%\Delta P$.....*Equation 4.5*

- ***Measurement***

The reader should be warned at this point that the following material will appear highly abstract at first, and for many people will only start to make sense when it is applied to the kinds of problem discussed. The definition is the easiest one to use for measurement. Referring to Figure 4.4 below,

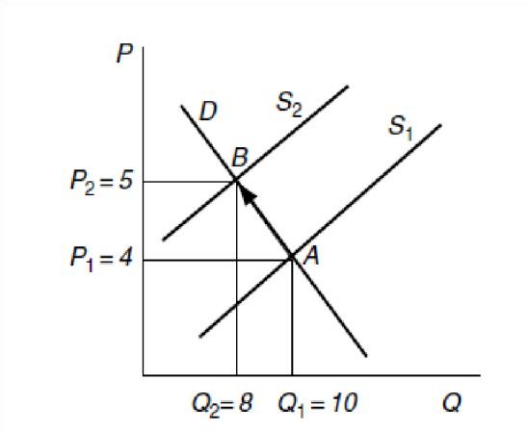


*Figure 4.4. Change in quantity demand* If price is raised from 4 to 5 units (the units do not matter):

PED from A to B = -20/ 25 = -0.8

The interpretation of this value is that for every 1 per cent rise in price there is a 0.8 per cent fall in the quantity demanded. However, there is a measurement problem here, because if we now reverse the price change and reduce the price to 4, the calculation is as follows:

PED from B to A = 25/ -20 = -1.25

Therefore, this measure of PED (termed the simple PED) is inconsistent; later on, it will also be shown to be inaccurate. The reason for this is the changing base for the percentage calculation. It is therefore better in terms of consistency and accuracy to use a measure involving average prices and quantities. Thus

$$\%\Delta Q = \frac{\Delta Q}{(\Delta Q_1 + Q_2)/2} = \frac{-2}{9} = -22.22$$

And

$$\%\Delta P = \frac{\Delta P}{(P_1 + P_2)/2} = \frac{1}{1.45} = 22.22$$

Therefore, in the above example PED = -1, both from A to B and from B to A; this measure of PED is termed the adjusted PED. The procedure above can be simplified by deriving the formula:

$$PED = \frac{\Delta Q}{\Delta P} \times \frac{P_1 + P_2}{Q_1 + Q_2}$$

$$= \frac{-2}{1} \times \frac{4+5}{10+8} = -1$$

It can easily be verified that the same result is obtained if the price change is from 5 to 4 units.

The above measures, both simple and adjusted, are called measures of arc elasticity, because they apply between two points or values. Sometimes we want to measure elasticity at a single point or price, and this is appropriate when we are considering the effect of a very small price change. This can be done using **point elasticity**, and involves the concept of limits and derivatives. The relevant formula becomes:

$$PED = \frac{\partial Q}{\partial P} + \frac{P}{Q} \quad \text{.....Equation 4.6}$$

Where $\partial Q/\partial P$ represents the partial derivative of Q with respect to P. This mathematical concept of a partial derivative has an important economic interpretation, which follows, the marginal effect of price on quantity, assuming other factors affecting demand remain the same. Normally this partial derivative is calculated by differentiating the demand function, in this case:

at A: $Q = 18 - 2P$

Therefore

at B: $PED = -2 \times 4/10 = -0.8$ and

$PED = -2 \times 5/8 = -1.25$

This shows that the PED varies along the linear demand curve.

*Table 4.3. Range of values for PED*

| \|PED\| | Demand | Interpretation |
|---|---|---|
| >1 | Elastic | Consumers responsive to price changes |
| <1 | Inelastic | Consumers not very responsive to price changes |
| =1 | Unit elastic | Intermediate case |
| ∞ | Perfectly elastic | Infinitely responsive (buy nothing if price rises) |
| 0 | Perfectly inelastic | Totally unresponsive (buy the same if price rises) |

With the power form of demand curve, we can again use either arc or point elasticity but in this case, we can prove that they are the same. If we apply the point PED formula to the power form $Q = aP^b$ we obtain the following from, using partial differentiation:

$$PED = baPb^{-1} \times \frac{P}{Q} = \frac{baP^{b-1}}{aP^b} \times \frac{P}{aP^b} = \frac{baP^{b-1}}{aP^b} = b \quad \text{.....Equation 4.7}$$

This is a very important result: it shows that for any power form of the demand curve the power represents the PED, which will be constant for all prices along the demand curve. This in turn means that the arc and point elasticity must be the same, as stated earlier. Furthermore, the result can be generalized to apply to any elasticity; for any power form the power of a variable represents the elasticity of the dependent variable with respect to that independent variable. This concept will be applied in many situations as we continue our analysis.

- ***Range of values***

This can be seen from Table 4.3. It should be noted that the PED is always interpreted in terms of its absolute magnitude when we are referring to whether demand is elastic or inelastic; its negativity is simply because of the inverse relationship between price and quantity demanded. We shall see, however, that in performing calculations involving the PED, it is important to include the sign.

- ***Determinants of PED***

There are three main factors that affect the price elasticity, of which the first is the most important.

*1. Availability of substitutes*

The closer a substitute(s) a product has the more elastic is its demand. An example is the demand for cigarettes: this demand is inelastic because such substitutes that exist (like cigars, pipes, chewing tobacco, other drugs, nicotine patches and chewing gum) are not very close in terms of their perceived functions and attributes. However, a single brand of cigarettes may have elastic demand, depending on brand loyalty. In this case we are referring to a different issue; instead of asking what will happen when the price of all cigarettes changes we are asking what will happen when the price of one brand changes while the price of other brands stays the same. In this situation the PED will depend on whether other brands are seen as substitutes; hence the importance of brand loyalty in affecting PED. If there is a great amount of brand loyalty, other brands will not be seen as substitutes and demand will be less elastic than would be the case if there were less brand loyalty. Generally speaking, the more narrowly a product is defined the more elastic its demand will be, for example food, breakfast and cereals have increasingly elastic demand.

*2. Proportion of income spent on the product*

The higher this is, the more elastic the demand, other things remaining equal. This is because the income effect of the price change is greater. Thus, the demand for sugar will tend to be inelastic while the demand for cars will be more elastic.

*3. Time frame*

Demand tends to be more elastic in the longer term, because it may take time for consumers to switch to different products. An example of this was oil in the 1970s, when the price quadrupled. The demand was inelastic in the short run because it was difficult for households and firms to cut back on their consumption. As time went on consumers could switch to other forms of heating and more fuel-efficient cars, and firms also switched to other firms of energy, such as electricity and gas-turbine generators. Habits are easier to change over a longer time period.

*Table 4.4. PED relationships with revenue, costs and profit*

| Demand | Price | Quantity | Revenue | Costs | Profit |
|---|---|---|---|---|---|
| Elastic | ↑ | ↓more | ↓ | ↓ | ? |
| | ↓ | ↑more | ↑ | ↑ | ? |
| Inelastic | ↑ | ↓less | ↑ | ↓ | ↑ |
| | ↓ | ↑less | ↓ | ↑ | ↓ |
| Unit elastic | ↑ | ↓same | same | ↓ | ↑ |
| | ↓ | ↑same | same | ↑ | ↓ |

↑ = increases; ↑more = increases by a larger proportion; ↑less = increases by a smaller proportion; ↑same = increases by the same proportion; ↓ = decreases; ↓more = decreases by a larger proportion; ↓less = decreases by a smaller proportion; ↓same = decreases by the same proportion; ? = unknown without further information

- ***Use and applications***

Firms want to know the PED for their products in order to charge the right price and to make forecasts. The ‗right ‘price is the price that achieves the firm ‘s objectives, for example profit maximization or revenue maximization. Some important relationships are shown in Table 4.4, which shows the effects of price changes on other variables, given different values of the

PED. At this stage the topic of costs has not been covered, but because the table is very general in terms of examining the directions of the changes rather than their precise sizes, it is sufficient to note that costs (meaning total costs) are directly related to output. It is

assumed here that this output corresponds to the quantity that people want to buy, implying that firms maintain fairly constant levels of inventory.

The above information can also be illustrated graphically. Figure 4.5, aids in understanding the relationships between the variables and the implications involved. Note that a linear demand curve is assumed, with an equation corresponding to the graph in Figure 4.5, this is given by Q = 18-2P, which readers should verify for themselves. Table 4.5 shows the relationships in Figure 4.5 in numerical form, ignoring costs. It should be noted that the marginal revenue can be calculated either from taking the differences in total revenue between each pair of prices or by taking the derivative of the revenue function; the latter method is used in Table 4.4. Table 4.5. Linear demand, PED and revenue *Figure 4.5. PED, revenue and profit.*

Tables 4.4 and 4.5, and Figure 4.5 show three things of particular importance:

*1. The arc elasticity using average prices and quantities is correct*

If revenue stays the same when there is a price change, demand must have unit elasticity; revenue at A and B in the earlier example was the same, at £40. This can be shown more formally by having a power form $Q = aP^{-1}$; this can be rewritten Q = a/P, where _a 'is a constant representing revenue.

2. *Firms will always maximize revenue by charging a price ($P_1$) where demand is unit elastic*

In this situation MR = 0. It can also be seen that the MR curve has twice the slope of the demand curve. Thus, the output $Q_1$ is halfway between the origin and point B, the horizontal intercept of the demand curve. This is proved as follows: The demand curve has equation P = a + bQ

Revenue, $R = PQ = (a + bQ)\ Q = aQ + bQ^2$

The MR curve has equation $MR = \underline{dR} = a + 2bQ$.....*Equation 4.8*

*3. Firms will always maximize profit by charging a price ($P_2$) where demand is Elastic* This can be explained in terms of Figure 4.5 by seeing that MC = MR (the condition for profit-maximizing output) at an output less than $Q_1$, since MR must be positive in this situation. Therefore, at this output ($Q_2$) the price must be higher than $P_1$, so demand must be in the elastic range. In order to determine the exact price that maximizes profit, more precise information regarding the demand and cost functions must be known.

As far as forecasts are concerned, the firm can estimate the effects on sales of a change in price if it knows the PED. In the above analysis we concentrated on the importance of the PED from the viewpoint of the management of the firm. Governments are also interested in price elasticities as a guide to various aspects of policy, for example the determination of levels of indirect taxes and import duties, and competition policy.

*4.1.1.2. Promotional (Advertising) Elasticity*

Promotional (Advertising) elasticity is now discussed; this is not because it is seen as being next in importance to price elasticity, since that position usually is taken by income elasticity, but because, like price, it is a controllable variable. Since, in practice, promotion often refers to advertising, the term _advertising elasticity 'will be used from now on.

***Definition***

Advertising elasticity (AED) can be defined in a very similar way to PED; AED refers to the percentage change in quantity demanded in response to a 1 per cent change in advertising. This means just changing the last word of the definition of PED to advertising (measured as expenditure). In symbols:

$AED = \%\Delta Q/\%\Delta A$ .....*Equation 4.9*

- ***Measurement***

As with PED both arc and point elasticities can be used. The simple measure of elasticity, using the formula above, should again not normally be used, unless the changes in advertising are very small, because it is inconsistent and inaccurate. The formulae for arc and point elasticities are similar to the formulae for PED, with the variable A substituted for P. Thus, we have the following:

Adjusted AED = $\frac{\Delta Q}{\Delta A} \times \frac{A1+A2}{Q1+2}$ ..... *Equation 4.10*

Point AED = $\frac{\partial Q}{\partial A} \times \frac{A}{Q}$ .....*Equation 4.11*

where $\partial Q/\partial A$ represents the partial derivative of Q with respect to A. This partial derivative has a similar economic interpretation to that involved in price elasticity; in this case it represents the marginal effect of advertising on quantity, assuming that other factors remain constant. With the power form of demand equation $Q = aP^bA^c$, it can be shown using partial differentiation that the advertising elasticity is given by c. We can see that all elasticities in a power form are given by the powers of the relevant variables and are constant; thus, the power form is sometimes known as the constant elasticity form.

- ***Range of values***

AED is expected to be positive, since we would expect consumers to react positively to promotion or advertising.

- ***Determinants***

The effectiveness of the firm ‘s advertising is important and so is the amount spent on advertising. As the firm spends more and more the market may become saturated (diminishing returns) and AED may decrease. There are no widely published results of empirical studies relating to promotional elasticities, and it would be difficult to generalize any conclusions even if there were.

***se and applications***

Firms want to know the AED for their products in order to spend the right amount on promotion and to make forecasts.

*4.1.1.3. Income elasticity*

- ***Definition***

Income elasticity (YED) is defined similarly to the above elasticities; it is the percentage change in quantity demanded in response to a 1 per cent change in income. Again, just the last word is changed, to _income ‘. In symbols:

$$YED = \%\Delta Q/\%\Delta Y \quad \text{.....Equation 4.12}$$

- ***Measurement***

This is also the same as for the above elasticities; there are three measures, simple, adjusted and point, but the first should not normally be used unless the changes in income are very small. Thus, the adjusted and point elasticities are:

$$\text{Adjusted YED} = \frac{\Delta Q}{\Delta Y} \times \frac{Y_1 + Y_2}{Q_1 + Q_2} \quad \text{.....Equation 4.13}$$

$$\text{Point YED} = \frac{\partial Q}{\partial Y} \times \frac{Y}{Q} \quad \text{.....Equation 4.14}$$

Where $\partial Q/\partial Y$ represents the partial derivative of Q with respect to Y. The power form of the demand equation would be extended to become $Q = aP^bA^cY^d$.

- ***Range of values***

Income elasticity can be positive or negative, depending on whether the product is normal or inferior. Normal products can be further divided into luxury products and staple products, according to whether the income elasticity is more than one or less than one respectively.

This is summarized in Table 4.6 below.

*Table 4.6. Range of values for YED*

| YED | Demand | Interpretation |
|---|---|---|
| > 1 | Income elastic | Luxury products |
| 0< YED< 1 | Income inelastic | Staple products |
| <0 | Negative elasticity | Inferior products |

***Determinants***

As can be seen above, the type of product is the main factor determining the income elasticity. The level of income in the market is also relevant; for example, products that are staple goods in developed countries, like telephones and radios, might be luxuries in a developing country. As with price elasticity, theory alone cannot tell us whether a product is in a certain category; we need empirical studies to confirm this.

- ***Use and applications***

Firms want to know the income elasticities for their products in order to select target markets and make forecasts. Note that income is not truly a controllable variable, but a firm can gain an element of control by selecting from different target markets with different levels of average income.

*4.1.1.4. Cross-elasticity*

- ***Definition***

This refers to the percentage change in the quantity demanded of one product in response to a 1 per cent change in the price of another product. This can be expressed in symbols:

$$CED = \%\Delta Q/\%\Delta P \text{ '} \quad \text{.....Equation 4.15}$$

- ***Measurement***

The same methods and principles apply as for the other elasticities. Thus:

$$\text{Adjusted CED} = \frac{\Delta Q}{\Delta P'} \times \frac{P'_1 + P'_2}{Q_1 + Q_2} \quad \text{.....Equation 4.16}$$

$$\text{Point CED} = \frac{\partial Q}{\partial P'} \times \frac{P'}{Q} \quad \text{.....Equation 4.17}$$

where $\partial Q/\partial$ Represents the partial derivative of Q with respect to P '. The power form of the demand equation would be extended to become $Q = aP^bA^cY^dP'^e$.

***Range of values***

CED can be positive or negative according to whether the other product is a substitute or a complement. For example, if the price of Pepsi rises, the quantity of Coke demanded should also rise, as people switch from Pepsi to Coke. Note the importance of the assumption of other things being equal, in particular that the price of Coke stays the same.

***Determinants***

As well as the above, the degree of substitutability or complementarity is important. The greater this is, the larger the absolute value of CED. Thus, in the example above, the closer Pepsi and Coke are seen as substitutes, the more people will switch from one to the other.

Again, it is important to test theoretical considerations by empirical studies.

***Use and applications***

Firms want to know the CED for their products in order to see what their main competition is and determine strategy accordingly; sometimes the ‗other products ‘involved are in the same product line or product mix of the firm itself. In this case, information relating to CEDs is important for product planning purposes. The information is again useful for making forecasts, when the price of a related product changes.

## 4.1.2. Demand Estimation

In the previous section it was generally assumed that the demand function for a firm or market was known; in practice it has to be estimated from empirical data, and that is the subject of this section. There are a variety of ways that can be used to estimate demand, each of which has certain advantages and disadvantages.

*4.1.2.1. Consumer surveys*

Firms can obtain information regarding their demand functions by using interviews and questionnaires, asking questions about buying habits, motives and intentions. These can be quick on-the-street interviews, or in-depth ones. They might ask, for example, how much more petrol respondents would buy if its price were reduced by 15 pence per litre, or which brand of several possibilities they prefer. These methods have certain drawbacks:

- ✓ *Validity*. Consumers often find it difficult to answer hypothetical questions, and sometimes they will deliberately mislead the interviewer to give the answer they think the interviewer wants.
- ✓ *Reliability*. It is difficult to collect precise quantitative data by such means.
- ✓ *Sample bias*. Those responding to questions may not be typical consumers.

In spite of these problems, there are advantages of surveys:

- ➢ They give up-to-date information reflecting the current business environment.

- Much useful information can be obtained that would be difficult to uncover in other ways;

For example, if consumers are ignorant of the relative prices of different brands, it may be concluded that they are not sensitive to price changes. Firms can also establish product characteristics that are important to the buyer, and priorities. Methods such as multidimensional scaling can be used to give rating scores on product characteristics.

*4.1. 2.2 Market experiments*

As with consumer surveys these can be performed in many ways. Laboratory experiments or consumer clinics seek to test consumer reactions to changes in variables in the demand function in a controlled environment. Consumers are normally given small amounts of money and allowed to choose how to spend this on different goods at prices that are varied by the investigator. However, such experiments have to be set up very carefully to obtain valid and reliable results; the knowledge of being in an artificial environment can affect consumer behaviour. Other types of market study involve using real markets in different geographic locations and varying the controllable factors affecting demand. This kind of test marketing has the advantage that direct observation of consumers 'actual spending behaviour is possible rather than just recording answers to hypothetical questions regarding such behaviour. There are still a number of problems with this method, however:

- There is less control in this case, and greater cost; furthermore, some customers who are lost at this stage may be difficult to recover.
- The number of variations in the variables is limited because of the limited number of market segments available. Thus, only a small number of sample observations is possible.
- Experiments may have to be long-lasting in order to reveal reliable indications of consumer behaviour.

We have seen in the previous section that price elasticity, for example, can be very different in the long run from in the short run because it takes time for consumers to change their habits.

*4.1.2.3 Statistical methods*

While the above methods are useful, they often do not provide management with the kind of detailed information necessary to estimate a useful demand function, and thereby test the relevant hypotheses and make forecasts. Statistical techniques, especially regression analysis,

provide the most powerful means of estimating demand functions. Regression techniques do have various limitations:

They require a lot of data in order to be performed.

- They necessitate a large amount of computation.
- They suffer from a number of technical problems,

In spite of these limitations, regression techniques have become the most popular method of demand estimation, since the widespread availability of powerful desktop PCs and software packages have made at least the first two problems easy to overcome. They are therefore the main subject of this section.

## 4.2. Pricing Techniques

Pricing is often treated as being the core of managerial economics. There is certainly a fair element of truth in this, since pricing brings together the theories of demand and costs that traditionally represent the main topics within the overall subject area. However, as indicated in various parts of this text, this can lead to an over-narrow view of what managerial economics is about. This section will continue to examine pricing in a broader context, but first it is helpful to consider the role of pricing in the economic system. Price determination is the core of microeconomics; this whole subject area examines how individual components in the economic system interact in markets of various kinds, and how the price system allocates scarce resources in the system.

Economists may well disagree about how well the market economy works in practice, and this aspect is discussed in the last chapter, but the general framework of analysis regarding the price system is not in serious dispute. Micro economists tend to focus on pricing at the level of the industry, and on the welfare aspects in terms of resource allocation. This perspective on price is what lies behind the assumption that price is the most important variable in the marketing mix. However, although this perspective is important for government policy, it is not the perspective that managers have, or should have, if their objective is profit or shareholder-wealth maximization, or indeed any of the other objectives discussed.

As far as managers are concerned, price is just one of many decision variables that they have to determine. The majority of managers do not consider it to be the most important of these decision variables, but they do tend to realize the interdependence involved in the decision-making process. Therefore, it makes sense

at this point to discuss the context of the pricing decision, before focusing on the more detailed analysis of pricing situations. The starting point for this discussion is the concept of competitive advantage.

## 4.2.1. Competitive advantage

The concept of competitive advantage was introduced by Porter in 1980, and has been utilized and further developed by many writers on business strategy since then. It provides a very useful means of analysing a firm ‘s success or lack of it in any market. A discussion of competitive advantage is therefore necessary in order to understand the nature of many of the decision variables that are involved in a firm ‘s business strategy, and to put these into a broad perspective.

As mentioned in the introduction to the section, this involves examining non-price decisions, and explaining the context within which price decisions are or should be made.

Nature of competitive advantage

In order to place the concept in context we initially need to recognize that a firm ‘s profitability depends in general on two factors: market conditions and competitive advantage.

**1. Market conditions**

These relate to external factors, not just for the firm, but also for the industry as a whole. Industries vary considerably in terms of their profitability; thus, throughout the 1990s, computer software firms, biotech firms and banks achieved higher than average profitability, while steel firms, coal producers and railways did badly. It is clear that these industry trends can vary from country to country, according to differences in external factors. These external factors are sometimes discussed in terms of another of Porter ‘s concepts, five-force analysis (internal rivalry, entry, substitutes and complements, supplier power and buyer power), or in terms of the value net concept of Branden burger and Nalebuff. A 1997 study by McGahan and Porter concluded that these factors explained about 19 percent of the variation in profit between firms.

**2. Competitive advantage**

This relates to internal factors, specifically those that determine a firm ‘s ability to create more value than its competitors. The study above concluded that these factors

explained about 32 percent of the variation in profit between firms. The concept of value creation now needs to be discussed.

- **Value creation**

The value created by a firm can be expressed in the following way:

Value created = perceived benefit to customer - cost of inputs

$$V = B - C$$

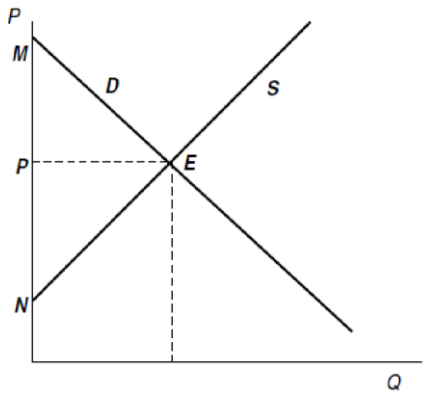


*Figure 4.17. Value, consumer surplus and producer surplus*

where V, B and C are expressed in monetary terms per unit. These concepts and their relationships to other important measures can be examined in terms of the graph in Figure 4.17. The demand curve shows the value that consumers place on buying an additional unit of a product, thus indicating perceived benefit in terms of marginal utility. The supply curve indicates the marginal cost to firms of producing an additional unit. Thus, value per unit is measured as the vertical distance between the demand and supply curves, and total value created is measured by the area MEN. The equilibrium price, P, at the intersection of the demand and supply curves can be seen to divide this value into two parts, consumer surplus (CS) and producer surplus (PS). Consumer surplus is given by B-P in general terms, which, as we have already seen in the last chapter, represents the amount that a consumer is prepared to pay over and above that which they have to pay. Producer surplus is given by P-C, which represents supernormal or economic profit. Thus, we can write:

$$CS = B - P$$

$$PS = P - C$$

$$V = CS - PS = (B - P) + (P - C) = B - C$$

For the first unit of consumption consumer surplus is given by M-P and producer surplus is given by P-N. At the equilibrium point no further value can be created by the firm by producing more output.

In order to make the situation less abstract, consider the following example.

A particular consumer may pay Birr 200 for a book; they may be prepared to pay Birr 250 for that particular edition, even though the marginal cost of production is only Birr 130. In this case the value created by that product item is Birr 120, of which Birr 50 is consumer surplus and Birr 70 is producer surplus.

In order to relate the above analysis to competitive advantage we need to introduce the concept of consumer surplus parity. We will assume to start with that consumers have identical tastes, though we will relax this assumption later. When we compare different products and making a purchase decision, rational consumers will try to maximize their total consumer surplus. Firms can be considered as making bids for consumers by offering them more surplus than competitors; a market will be in equilibrium when each product is offering the same surplus as its rivals. Thus in the athletic shoe market a particular model of Nike shoe may sell for Birr1000 more than a competitor ‘s product, but if consumers perceive the benefit to be Birr1000 greater, they will obtain the same surplus from consuming each product and therefore be indifferent between the two products. This equilibrium will mean that each product would maintain its market share relative to other products.

If a firm (or product) has a competitive advantage, it is in a position to make more profit than competitors. Firms are able to do this essentially in two different ways: pursuing a cost advantage or pursuing a benefit advantage. These concepts are examined in more detail shortly but can be briefly summarized here:

1. ***Cost advantage***. In this case the greater value created (B-C) depends on the firm being able to achieve a lower level of costs than competitors, while maintaining a similar perceived benefit.
2. ***Benefit advantage***. In this case the greater value created depends on the firm being able to achieve a higher perceived benefit than competitors, while maintaining a similar level of costs.

In order to discuss these issues further we must now consider the assumptions on market structure, regarding the nature of the pricing decision. These included, in particular:

1. The firm produces for a single market.
2. The firm charges a single price for the product throughout the whole market.
3. The firm produces a single product rather than a product line.
4. The firm produces a single product rather than a product mix.
5. The firm is considered as a single entity, rather than as consisting of different divisions, sometimes in different countries.
6. Price is the only instrument in the marketing mix.
7. Pricing is considered from a static, single-period viewpoint.

These assumptions are relaxed in the following sections, taking each in turn. The following case study applies the concept of competitive advantage to a very high-profile firm in the telecommunications industry.

### 4.2.2. Market positioning, segmentation and targeting

Positioning in the market is the most fundamental aspect of a firm ‘s marketing strategy. A firm must begin by examining its resources and capabilities relative to the business that it is in, or any business that it could be in. These resources and capabilities include things like management know-how, technical expertise, reputation and image, patents and trademarks, organizational culture, ownership of specialist resources, relationships with customers and suppliers and location. It should be noted that these factors tend to be ingrained within the organization, meaning that they are not dependent on the presence of certain individuals. Another important characteristic is that they are not easily duplicated. If there is a close match between these firm specific factors and the key success factors in the industry then the firm may be able to obtain a competitive advantage in either cost or benefit terms.

*4.2.2.1. Cost advantage*

In basic terms this means achieving a lower level of costs while maintaining a level of benefits that is comparable to competitors. In some cases, the firm claims to have the same quality as competitors at a lower cost, while in other cases the firm ‘s perceived quality may be lower, but the cost and price substantially lower. In this situation we no longer need to assume that consumers have identical tastes; the firm can provide at least as much consumer surplus to some customers as competitors, while maintaining at least the same profit. The determination of how the value

created should be divided into consumer surplus and producer surplus involves pricing strategy. Cost advantage can be achieved in a number of ways: economies of scale and scope, the learning curve, production efficiency, control over inputs, transaction efficiencies and favourable government policies may all play a part.

We will take the car industry as an overall example. In the case study on Nissan, it was seen that transactional efficiencies in terms of dealing with suppliers formed an important component of a cost reduction strategy, as did increasing capacity utilization. When BMW took over Rover in the UK they were hoping, among other things, to obtain favourable government treatment in terms of grants and subsidies. Other car firms are concentrating on producing a smaller number of basic platforms for their different models in order to gain more economies of scale. Korean car manufacturers, like Hyundai and Daewoo, have concentrated on producing no-frills cars of lower perceived quality, but at a much lower price than other cars. The British sports car manufacturer TVR produces high-performance cars at a lower price than cars of similar performance levels produced by other manufacturers; this is achieved by simplifying the production process, using components from other cars, omitting some high-tech gadgetry and cutting back on marketing overheads.

*4.2.2.2. Benefit advantage*

A strategy aimed at achieving a benefit advantage involves offering a product with a higher level of benefit than competitors while maintaining a similar level of cost.

Examples of firms in the car industry that have pursued this type of strategy are Honda, BMW and Toyota with its Lexus model, although with this Toyota originally pursued both ends of the spectrum by producing a car at a lower price than the luxurious Mercedes and BMW saloons, but which used and provided more in the way of new technology. Other companies aiming at a benefit advantage may price their products somewhat higher than competitors but boast a significantly greater quality. Porsche is an example. Some companies, like Aston Martin and Ferrari, take this to the extreme; they charge very high prices, but promise the very best in quality. Just as with cost advantage, benefit advantage can be achieved in a number of ways, even in a given industry. Reputation or marque counts for a large amount in the luxury car market. The Japanese manufacturers in general have tried to make up for a lack of this characteristic by using and providing more advanced

technology, such as ABS, traction control, variable valve timing, four-wheel steering, climate control, satellite navigation and other gadgets. Aston Martin emphasizes build quality and customer service. BMW stresses German 'engineering and reliability. It should be emphasized however that firms pursuing a benefit advantage do not have to be producing a luxury product; the quality of the product does not have to be high in general, it simply has to be perceived as being higher than competitors in that particular market segment. For example, the Japanese manufacturers mentioned earlier have mostly been producing cars in the medium or low-price range. Again, as with cost advantage, the firm must determine how the value created should be divided into consumer and producer surplus. This means determining a pricing strategy. The general considerations regarding this are discussed in the next subsection.

### 4.2.3. Competitive advantage, price elasticity and pricing strategy

We now need to apply the concept of price elasticity of demand (PED), examined to the concept of competitive advantage in order to see how firms should generally price their products.

**a. Cost advantage**

There are two possible strategies for firms with a cost advantage:

1. The greater value that it creates can be given over to customers in the form of consumer surplus by charging a lower price, with the firm maintaining the same profit margin as competitors; in this case the firm should increase its market share and thus make more total profit.
2. The greater value created can be translated into producer surplus, meaning a greater profit margin, by charging the same price as competitors and providing the same perceived quality.

*Table 4.12. Price elasticity and competitive advantage*

| | High PED | Low PED |
|---|---|---|
| Cost advantage | Cut price<br>Gain market share | Maintain price<br>Gain profit margin |
| Benefit advantage | Maintain price<br>Gain market share | Increase price<br>Gain profit margin |

Choice of the appropriate strategy depends on the PED of the product. If PED is high, a lower price will allow the firm to capture significant market share from competitors. On the other hand, if PED is low, large price cuts will not increase market share much; in this case it is better to keep price at the same level as competitors and increase profit margin; see Table 4.12.

**b. Benefit advantage**

Again, there are two possible strategies:

1. The greater value created can be given over to customers in the form of a larger consumer surplus by charging the same price as competitors and translating the greater consumer surplus offered into greater market share.

2. The greater value created can be translated into producer surplus by charging a higher price than competitors, while still maintaining the same consumer surplus. This will increase profit margin. In this situation of benefit advantage, if PED is high it is better to maintain price at the same level as competitors and gain market share. If PED is low then it is better to increase price, since this will not affect market share much, and concentrate on increasing profit margin; again, see Table 4.12.

### 4.2.4. Segmentation and targeting

These two aspects of strategy are closely related to competitive advantage and positioning. A market segment refers to a group of consumers in a wider market who are homogeneous in some respect related to their buying behaviour.

Markets can be segmented in many ways according to different characteristics: demographic factors (income, age, ethnicity and sex, for example), location, and frequency of purchase or use, lifestyle and newspaper readership are all common means of segmentation.

Why is it useful for firms to segment markets? Different segments have different tastes and therefore respond differently to the various marketing mix variables. A high-income segment, for example, may have a lower PED for a product than a lower income segment, meaning that they may be less price sensitive; on the other hand, they may be more quality-sensitive. A firm may therefore want to select target markets according to its competitive advantage. However, strategy selection does not always occur in this order. For example, a firm may observe that a certain market segment is not currently being adequately catered for in the market; it may therefore pursue a competitive advantage in producing some product or range of products that appeal to that segment. In the first case we could say that the firm is supply driven,

while in the second case the firm is demand-driven. Targeting strategies are often classified into two main categories: broad coverage strategies and focus strategies.

**a. Broad coverage strategies**

These strategies aim to provide for the needs of all or most of the segments in the market. This can be achieved either by providing a general product line of a _one-size-fits-all ‘type, or by providing a product line consisting of a variety of related products that are more-or-less customized for each segment. An example of the first approach is the strategy used by most commercial banks for attracting bank depositors, at least until recently. Little distinction was made according to age, income, occupation or frequency of cheque-writing. The main attraction of this approach is that it allows more standardization, and therefore leads to greater economies of scale. However, most firms try to appeal to different segments by producing different products in a product line.

**b. Focus strategies**

These involve a firm either offering a single product, or serving a single market segment, or both. The first strategy may apply in markets where there is a lot of product homogeneity, like steel production, or it may apply to small specialist firms, particularly in industrial markets. More commonly a firm will specialize in providing for a single market segment; for example, TVR specializes in producing high performance sports cars, as do Lotus and Marcos. Bose specializes in producing high-quality speaker systems. Tag Heuer specializes in producing high-quality watches, particularly for sporting people. However, the focus does not have to be on high quality, associated with benefit advantage. Firms like Asda, K-mart, McDonald ‘s and Woolworth have specialized in providing goods of adequate ‘quality at a low price. In this case, economies of scale and other cost reduction factors are of vital importance in yielding a cost advantage.

## 4.2.5. Role of pricing in managerial decision-making

We can now see that pricing is best considered as one of many interdependent decisions that firms have to make, and often is not the most fundamental. Firms need to start from their objectives in general terms, and then consider their competitive advantage in terms of their resources and capabilities, in conjunction with the external conditions of the market in terms of customers and competitors.

Normally this means starting with a positioning decision, leading to a product decision, and then considering the other elements of the marketing mix. Pricing is considered in this context, and affects the buying decisions of consumers through determining the amount of consumer surplus created.

Firms can pursue either a cost advantage or a benefit advantage. In either case the firm has to create more value (Benefit - Cost) than competitors. A firm then must determine a pricing policy to divide this value between consumer surplus and producer surplus. If a firm wants to increase market share, it has to create more consumer surplus (Benefit-Price) than its competitors. An alternative strategy is for the firm to increase profit margin by creating more producer surplus (Price-Cost).

It can be seen, therefore, that the main priority of management is to identify and determine products where value creation is possible. Only then is the division of such value into consumer and producer surplus considered. Thus, pricing is usually seen as being a secondary consideration.

## 4.3. Price discrimination

### a. Definition

Price discrimination has been defined in a number of different ways. The simplest definition relates to the situation where a firm sells the same product at different prices to different customers. However, the most useful definition involves a firm selling the same or similar products at different prices in different markets, where such price differentials are not based on differences in marginal cost. This latter definition covers a broader range of situations and leads to a greater understanding regarding the nature of price discrimination.

Using this definition, we can consider the following common practices as price discrimination:

* Airlines that charge different fares for the same journey at different seasons of the year.
* Universities that charge higher fees to overseas students than to home students.
* Restaurants that offer _early bird ‘dinners.
* Professional people, like doctors, accountants and consultants, who charge richer clients higher fees than poorer clients for the same service.
* Exporters who charge lower prices abroad than they charge domestically.
* Prestige outlets that charge higher prices than high street or discount stores for the same products.

* Health clubs that sell off-peak memberships.
* Supermarkets that offer points or reward schemes to regular shoppers.
* Happy hour.

We can now consider the conditions that are necessary for price discrimination to be possible.

**b. Conditions**

There are two fundamental conditions:

1. There must be different market segments with different demand elasticities.
2. The market segments must be separated so that there is no possibility of resale from one segment to another.

In order to illustrate the importance and application of these conditions let us consider the example of a butcher. Instead of advertising his meat prices he may wait until the customer comes into the shop and charge prices according to an assessment of the individual customer ‘s ability and willingness to pay. Such a strategy may be successful in the case of customers who are ignorant of the prices being charged to other customers; however, once they become aware of this situation, and assuming that other butchers are following the same practice, high-paying consumers will find ways to avoid paying higher prices by buying their meat from other customers, in other words arbitrage will occur. Arbitrage refers to the practice of buying at a lower price and selling at a higher price to make a riskless profit. Thus, in this instance a butcher cannot successfully practise price discrimination. On the other hand, a dentist is in a position to do so; patients cannot ask other patients to have their teeth capped for them. Thus, price discrimination is generally easier in the markets for services, especially personal and professional services. At this point one might ask why a firm would want to practise price discrimination. Essentially it is a method of reducing the amount of consumer surplus and transferring it to the producer in terms of additional profit.

This aspect is most easily seen by considering different types of price discrimination.

## 4.3.1. Types of price discrimination

There are different ways of classifying these types of discrimination. We will discuss the degree of price discrimination first, since it leads to a better understanding of the resulting impact on profit; then bases for price discrimination will be considered, in terms of different ways of segmenting markets.

**a. Degree of price discrimination**

Economists tend to classify price discrimination into three main categories, according to the extent to which consumer surplus is transferred to the producer.

1. **First-degree price discrimination.** This is sometimes referred to as perfect price discrimination, since all the consumer surplus is transferred to the producer. For this to be possible the producer must have complete knowledge of its demand curve, in terms of knowing the highest price that each customer is prepared to pay for each unit, and be able to sell the product accordingly. This situation is extremely rare, but an example is an auction market, like the Treasury Bill market. Figure 4.18 illustrates the situation. It is assumed in this case that marginal costs are constant, for simplicity. The whole consumer surplus, $GCP_3$, becomes revenue and profit for the producer.

2. **Second-degree price discrimination**. In this situation a firm may charge different prices for different levels of consumption. The first $Q_1$ units may be sold at the high price of $P_1$; the next block of units, $Q_2\_Q_1$, may be sold at the price $P_2$, and the last block of units, $Q_3\_Q_2$, may be sold at the price of $P_3$. This situation is also shown in Figure 4.18 and the size of the consumer surplus is given by the three triangles, $GAP_1$, ABE and BCF. This type of strategy may be used in the selling of a product like electricity. In other situations, a reverse type of strategy is sometimes used, where early buyers pay a lower price and later buyers have to pay more. Examples are the selling of tickets to certain events, like concerts, and the selling of certain types of club membership. Usually buying early involves more inconvenience or inflexibility for the buyer, thus segmenting the market. Second-degree price discrimination often features a two-part tariff or pricing system. There is some lump sum charge and then a user charge.

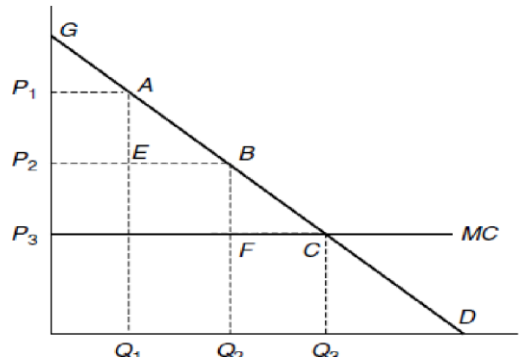

*Figure 4.18. Degrees of price discrimination*

**3. Third-degree price discrimination**. This is, in practice, the most common form, where a firm divides its market into segments and charges different prices accordingly. Markets can be segmented in various ways for this purpose, and this aspect is now discussed. **b. Segmentation for price discrimination**

Firms can segment markets on different bases; the following are the most common.

**1. Time.** Many products that have seasonal demand have different prices at different time periods; airlines, hotels, restaurants, cinemas, power providers and many other industries have this feature. Demand is less elastic at peak season, so prices are higher then.

**2. Location**. In some cases, the price of a product may be higher in one location than another because of a higher transportation or other cost. This is not price discrimination. However, when the price in one location is higher than is justified by a difference in marginal cost, because of demand being less elastic, then this does involve discrimination. It should also be mentioned that sometimes a kind of reverse price discrimination occurs. Instead of charging different prices when marginal cost is the same, some firms will charge the same price even when marginal costs are different. This occurs, for example, when there are differences in transportation costs between different markets, but a universal price is charged. Again, this could be justified by differences in demand elasticities between different markets, if demand is more elastic in those areas with higher transportation costs.

**3. Product form**. Many firms offer a product line where the top-of-the-line model is considerably more expensive than the other products in the line. This may be true even when there is little difference in the cost of production of the different products. In this case there may be some consumers who are determined to have the best (stereo, television, car, mountain bike, tennis racquet), and therefore, again, their demand is less elastic in the normal price range.

**4. Customer.** This is the most controversial type of price discrimination because of concerns regarding fairness. Unfortunately, the term _discrimination ‘is an emotive word and tends to be associated with normative issues, but it is not our purpose to discuss these here. It is simply necessary to point out that firms can increase their profit by charging higher prices to customers whose demand is less elastic, provided that there is no possibility of resale between market segments, as

explained earlier. This explains the discrimination carried out by many providers of professional services, including universities. It also explains the price discrimination practised by many exporters, charging higher prices in the domestic market in spite of lower costs.

## 4.4. Other pricing strategies

This chapter has so far focused on the pricing behavior of profit-maximizing firms operating under somewhat unique circumstances. In each case, the firm ‘s pricing practices were predicated on subtle economic concepts. It was also assumed that management had complete information about the realities of the market in which the firm operated. In practice however, a firm ‘s pricing practices are much looser in the sense that they are based less on detailed mathematical analysis than on perception, custom, and intuition. There are five of these alternative pricing practices-*price leadership*, *price skimming*, *penetration pricing*, *prestige pricing*, and *psychological pricing*.

- **Price leadership**

Price leadership is a phenomenon that is likely to be observed in oligopolistic industries. Oligopolistic industries are characterized by the interdependence of managerial decisions between and among the firms in the industry. Firms in oligopolistic industries are keenly aware that the pricing and output decisions of any individual firm will provoke a reaction by competing firms. A consequence of this interdependence is relatively infrequent price changes. Price leadership occurs when a dominant company in an industry establishes the selling price of a product for the rest of the firms in the industry.

- **Price skimming**

If a firm is first to market with a new product, it may engage in a form of first-degree price discrimination called price skimming. As with first degree price discrimination, price skimming is an attempt to extract consumer surplus. During the interval between the firm ‘s introduction of a new product and the competition ‘s development of their own versions of the new product, the innovating firm is a virtual monopoly. If the innovating firm wants to extract consumer surplus, however, it must act fast.

Price skimming is an attempt by a firm that introduces a new product to extract consumer surplus through differential pricing before competitors develop their own versions of the new product. The firm begins by initially charging a very high price for its product. Consumers willing and able to pay this price will buy first. Before competitors have a chance to sell their versions of the new product, the innovating firm will lower its price just a bit to attract the next, lower tier of consumers. This process is continued until the price charged equals the marginal cost of production. Pricing in this manner will enable the innovating firm to extract consumer surplus to enhance profits. Of course, for this pricing scheme to be successful the firm must have knowledge of the demand curve of the product. Management is unlikely to have such knowledge, however, because of the novel nature of the product. The firm could, of course, conduct consumer surveys, market experiments, and so on, to develop information regarding the demand for the product, but care would have to be taken not to ―tip off‖ the competition.

- **Penetration pricing**

Penetration pricing occurs when a firm entering a new market charges a price that is below the prevailing market price to gain a foothold in the industry. A form of penetration pricing is *dumping*. Dumping is defined by the U.S. Department of Commerce as selling a product at less than fair market value. The most egregious form of this kind of pricing behaviour is *predatory dumping*, the attempt by a foreign producer to gain control of the market in another country by selling a product there at less than fair market value with the goal of driving out domestic producers. Penetration pricing is the practice of charging a price that is lower than the prevailing market price to gain a foothold in the industry.

- **Prestige pricing**

Prestige pricing is essentially an attempt by firms to increase sales of certain products by capitalizing on snob appeal. Many consumers derive a degree of personal identity from the ostentatious display of certain brand name items. For them, an enhanced personal image from the conspicuous consumption of upscale products is as valuable, and sometimes more so, than the usefulness or quality of the product itself. The mere fact that a product sells for a higher price often conveys the impression of higher quality, which may or may not be supported by reality. Prestige pricing is an attempt by some firms to exploit this perception by charging

higher prices because of the increased prestige that they believe ownership of their products confers. An often-cited example is the luxury automobile market, where higher priced automobiles are perceived to be superior to lower priced automobiles of similar quality.

Prestige pricing is the practice of charging a higher price for a product to exploit the belief by some consumers that a higher price means better quality, which in turn confers on the owner greater prestige.

- **Psychological pricing**

Finally, psychological pricing is a marketing ploy designed to create the illusion in the mind of the consumer that a product is being sold at a significantly lower price when, in fact, the price differential is inconsequential. Retailer sale of a product for $4.99 instead of $5.00 is psychological pricing. Retailers who engage in psychological pricing are attempting to exploit consumers 'initial impressions or their lack of familiarity with the product. The effects of psychological pricing tend to be transitory, however, as initial impressions wear off or the consumer becomes more knowledgeable about the product.

Psychological pricing is a marketing ploy designed to create the illusion in the mind of the consumer that a product is being sold at a significantly lower price when, in fact, the price differential is inconsequential.

***Self Assessment Exercise***

1. The X-Corporation produces a good (called *X)* that is a normal good. Its competitor, Y Corp., makes a substitute good that it markets under the name —*Y."* Good *Y* is an inferior good.
    a. How will the demand for good *X* change if consumer incomes increase?
    b. How will the demand for good *Y* change if consumer incomes decrease?
    c. How will the demand for good *X* change if the price of good *Y* decreases?
    d. Is good *Y* a lower-quality product than good *X?* Explain.
2. Suppose the own price elasticity of demand for good *X* is -2, its income elasticity is 3, its advertising elasticity is 4, and the cross-price elasticity of demand between it and good *Y* is -6. Determine how much the consumption of this good will change if:
    a. The price of good *X* increases by 5 percent.
    b. The price of good *Y* increases by 10 percent.

c. Advertising decreases by 2 percent.
d. Income falls by 3 percent.

3. If Starbucks ‘s marketing department estimates the income elasticity of demand for its coffee to be 1.75, how will looming fears of a recession (expected to decrease consumers ‘incomes by 4 percent over the next year) impact the quantity of coffee Starbucks expects to sell?
4. You are a division manager at Toyota. If your marketing department estimates that the semi-annual demand for the Highlander is $Q = 100{,}000 — 1.25P$, what price should you charge in order to maximize revenues from sales of the Highlander?
5. For the first time in two years, Big G (the cereal division of General Mills) raised cereal prices by 2 percent. If, as a result of this price increase, the volume of all cereal sold by Big G dropped by 3 percent, what can you infer about the own price elasticity of demand for Big G cereal? Can you predict whether revenues on sales of its Lucky Charms brand increased or decreased? Explain.

# CHAPTER FIVE
# 5. BUSINESS STRATEGY AND GAME THEORY

***Objectives***

- *To elucidate the nature and scope of game theory*
- *To show how managerial decisions can be supported by different game theories and business strategy*
- *To explain different models of oligopoly and their importance for managerial decisions*

## 5.1. Nature and scope of game theory

The essential nature of game theory is that it involves strategic behaviour, which means interdependent decision-making. We have at this point seen quite a few examples of situations where such decision-making is involved, particularly in the areas of the theory of the firm and competition theory. Some examples which will be analysed in more detail concern the tragedy of the commons, contracting between firms, contracting between an employer and employee, and oligopolistic situations in terms of determining price and output, and limiting entry. The essence of these interdependent decision-making situations is that when A decides (for example regarding price, entry into a market, whether to take a job), it will consider the reactions of other persons or firms to its different strategies, usually assuming that they act rationally, and how these reactions will affect their own utility or profit. It must also consider that the other parties (from now on called players), in selecting their reactive strategies, will consider how A will react to their reactions. This can continue in a virtually infinite progression.

In this situation there is often a considerable amount of uncertainty regarding the results of any decision. These kinds of situation occur in all areas of economics, not just in managerial economics. They are common in macroeconomic policy, labour economics, financial economics and international economics. Game theory situations also occur in politics, sociology, warfare, _games ‘and sports, and biology, which make the area a unifying theme in much analysis. Game theorists, therefore, have come from many different walks of life, although the main pioneers were von Neumann and Morgenstern and Nash who were essentially mathematicians.

**Elements of a game**

The concept of a game, as we are now using it, embraces a large variety of situations that we do not normally refer to as games. Yes, chess, poker and rock–paper–scissors are games in the conventional sense, as are tennis and football (either American football or soccer); but games also include activities like going for a job interview, a firm bargaining with a labour union, someone applying for life insurance, a firm deciding to enter a new market, a politician announcing a new education/transport/health policy, or a country declaring war. *Table 5.1. Prisoner's Dilemma*

| | | Suspect B: Confess | Suspect B: Not Confess |
|---|---|---|---|
| Suspect A: Confess | | 5 / 5 | 10 / 0 |
| Suspect A: Not Confess | | 0 / 10 | 1 / 1 |

What do these diverse activities have in common? The following are the key elements of any game:

1. Players. These are the relevant decision-making identities; whose utilities are interdependent. They may be individuals, firms, teams, social organizations, political parties or governments.
2. Strategies. These are complete plans of action for playing the game. Although strategies may simply involve choosing a single action, it is important to understand that in some games there may be many actions involved. A complete plan means that every possible contingency must be allowed for.
3. Payoffs. These represent changes in welfare or utility at the end of the game, and are determined by the choices of strategy of each player. It is normally assumed that players are rational and have the objective of maximizing these utilities or expected utilities. Notice that the word each is important; what distinguishes game theory from decision theory is that, in the latter, outcomes only depend on the decisions of a single decision-maker.

The normal-form representation of a game specifies the above three elements. Now that we have stated the key elements in games, an example of a game can be presented, which will aid the ensuing discussion. The example is the well-known Prisoner ‘s Dilemma (PD). The reasons for presenting this situation here are twofold. First, although it does not directly involve a business predicament, the situation can easily be changed to a business predicament, and this is done shortly. Second, the conclusions regarding strategy appear paradoxical; this is a common finding in game theory, many conclusions being counter-intuitive. The classic PD situation involves two prisoners who are held in separate police cells, accused of committing a crime. They cannot communicate with each other, so each does not know how the other is acting. If neither confesses, the prosecutor can get them convicted only on other minor offences, each prisoner receiving a one-year sentence. If one confesses while the other does not, the one confessing will be freed while the other one receives a ten-year sentence. If both confess they will each receive a five-year sentence. This game is represented in normal form in Table 5.1.

*Table 5.2. Prisoner’s Dilemma for Coke and Pepsi*

| | | Pepsi | |
|---|---|---|---|
| | | Maintain Price | Discount |
| Coke | Maintain Price | 50 / 50 | 70 / –10 |
| | Discount | –10 / 70 | 10 / 10 |

The values in the table represent payoffs, in terms of jail sentences; the payoffs for Suspect A are below the diagonal, while the payoffs for Suspect B are above the diagonal. The objective for each suspect in this case is obviously to minimize the payoff in terms of jail time. The problem that they have in this case is that the best combination payoff for the pair of them is for them both not to confess, in other words to _co-operate ‘with each other. However, as we shall see shortly, this is not an equilibrium strategy. The equilibrium strategy is for both suspects to confess, or to _defect ‘. This equilibrium situation represents a paradox, since they will both end up serving a longer time in jail than if they had co-operated. The equilibrium still applies even if the suspects had agreed to co-operate beforehand; they will still

tend to defect once they are separated and do not know how the other is acting. The reader may wonder at this stage what the type of situation described above has to do with business strategy.

To illustrate this, let us consider the situation of Coke and Pepsi. At any given time period each firm has to decide whether to maintain their existing price or to offer a discount to the retailers who buy from them. Table 5.2 shows the payoff matrix for this situation, with the payoffs referring to profits, measured in millions of dollars per month. The payoffs for Pepsi are shown above the diagonal and the payoffs for Coke are shown below it. The objective in this situation is to maximize, rather than minimize, the payoffs. The reader may notice an important difference between the situation described in the original Prisoner ‘s Dilemma (PD) in Table 5.1 and that involving Coke and Pepsi in Table 5.2. The first situation is a _one-off ‘whereas the second is likely to be repeated regularly. As we shall see, this makes a big difference in determining optimal strategy.

In this game, maintaining price represents co-operating, while offering a discount represents defecting. As with the original Prisoner ‘s Dilemma in Table 5.1, there is a specific ordering of payoffs, as shown in Table5.3. Just as in the Prisoner ‘s Dilemma, the best combined payoff, or profit, is if both firms co-operate and maintain price. Note that co-operation here does not mean explicit collusion; rather it refers to tacit co-operation. Again, this is not an equilibrium situation; the equilibrium (explained shortly) is for both firms to discount.

*Table 5.3. Structure of payoffs in Prisoner’s Dilemma*

| Strategy pair (self/other) | Name of payoff |
|---|---|
| Defect/co-operate | Temptation (70) |
| Co-operate/co-operate | Reward (50) |
| Defect/defect | Punishment (10) |
| Co-operate/defect | Sucker's payoff (−10) |

Before we move on to an analysis of this situation and explain the conclusions above, it is helpful to broaden the perspective by considering the different types of game that can occur in business situations.

## 5.2. Types of game

There are many different types of game theory situation, and different methods of analysis are appropriate in different cases. It is therefore useful to classify games according to certain important characteristics.

**a. Co-operative and non-cooperative games**

In co-operative games the players can communicate with each other and collude. They can also enter into third-party enforceable binding contracts. Much of this type of activity is expressly prohibited by law in developed countries, so most of the games that are of interest in economic situations are of the non-cooperative kind. This type of game involves forming self enforcing reliance relationships, which determine an equilibrium situation.

**b. Two-player and multi-player games**

Both versions of the PD situation above are two-player games. However, both games are capable of being extended to include more than two parties. This tends to increase the likelihood of defection, particularly in the _one-off 'situation. Such a situation is sometimes referred to as _the tragedy of the commons '. The reasoning is that with more players it is important to defect before others do; only if defectors are easily detected and punished will this be prevented. Again, this has important implications for oligopoly; it is also relevant in international relations. The depletion of fish stocks in the North Sea due to overfishing, and the resulting conflicts, are an example of the tragedy of the commons. Property rights theory is obviously relevant in this area. With multi-player games there is also the opportunity for some of the players to form coalitions against others, to try and impose strategies that would otherwise be unsustainable.

**c. Zero-sum and non-zero-sum games**

With zero-sum games the gain of one player(s) is automatically the loss of another player(s). This can apply for example in derivatives markets, where certain transactions occur between two speculators. However, most business situations involve non-zero-sum games, as can be seen in the Coke/Pepsi situation earlier. The combined profit of the two firms varies according to the strategies of both players.

**d. Perfect and imperfect information**

In both versions of the PD presented earlier it was assumed that all the players knew for certain what all the payoffs were for each pair of strategies. In practice this is

often not the case, and this can also affect strategy. In some cases, a player may be uncertain regarding their own payoffs; in other cases, they may know their own payoffs but be uncertain regarding the payoffs of the other player(s). For example, an insurance company may not know all the relevant details regarding the person applying for insurance, a situation leading to adverse selection. Likewise, bidders at an auction may not know the valuations that other parties place on the auctioned item. Games with imperfect information are, unsurprisingly, more difficult to analyse.

**e. Static and dynamic games**

Static games involve simultaneous moves; the PD game is a simultaneous game, meaning that the players make their moves simultaneously, without knowing the move of the other player. In terms of analysis the moves do not have to be simultaneous in chronological terms, as long as each player is ignorant of the moves of the other player(s). Many business scenarios involve dynamic games; these involve sequential moves, where one player moves first and the other player moves afterwards, knowing the move of the first player. Investing in building a new plant is an example of this situation. As we shall see, the order of play can make a big difference to the outcome in such situations.

**f. Discrete and continuous strategies**

Discrete strategies involve situations where each action can be chosen from a limited number of alternatives. In the PD game there are only two choices for each player, to confess or not confess; thus, this is a discrete strategy situation. In contrast, a firm in oligopoly may have a virtually limitless number of prices that it can charge; this is an example of a continuous strategy situation. As a result, the analytical approach is somewhat different, as will be seen in the subsection on oligopoly.

**g. 'One-off' and repetitive games**

The distinction between these two types of situation has already been illustrated by the two different versions of the PD in Tables 5.1 and 5.2. Most short run decision scenarios in business, such as pricing and advertising, are of the repetitive type, in that there is a continuous interaction between competitors, who can change their decision variables at regular intervals. Some of these games may involve a finite number of plays, where an end of the game can be foreseen, while others may seem infinite. Long-run decisions, such as investment decisions, may resemble the _one-

off ‘situation; although the situation may be repeated in the future, the time interval between decisions may be several years, and the next decision scenario may involve quite different payoffs.

## 5.3. Static games

As stated above, these games involve simultaneous moves, as in the PD game. We shall concern ourselves at this stage only with games involving perfect information, and with _one-off ‘situations; all the players know with certainty all the possible payoffs, and the game is not repeated. The nature of this type of game raises the following questions:

1. How does a firm determine strategy in this type of situation?
2. What do we mean by an equilibrium strategy?
3. Is there anything that firms can do to change the equilibrium to a more favourable one, meaning to ensure co-operation?

These issues are addressed in the following subsections.

**Equilibrium**

In order to determine strategy or an equilibrium situation, we must first assume that the players are rational utility maximizers. We can now consider three types of equilibrium and appropriate strategies in situations involving different payoffs. These are dominant strategy equilibrium, iterated dominant strategy equilibrium, and Nash equilibrium. It is important to consider these equilibria in this order, as will be seen.

**a. Dominant strategy equilibrium**

A strategy $S_1$ is said to strictly dominate another strategy $S_2$ if, given any collection of strategies that could be played by the other players, playing $S_1$ results in a strictly higher payoff for that player than does playing $S_2$. Thus, we can say that if player A has a strictly dominant strategy in a situation, it will always give at least as high a payoff as any other strategy, whatever player B does. A rational player will always adopt a dominant strategy if one is available.

Therefore, in any static game involving discrete strategies, we should always start by looking for a dominant strategy. This is easiest in a two-strategy situation; when there are many possible strategies, dominant strategies have to be found by a process of eliminating dominated strategies, as shown in the subsection on Nash bargaining.

To start with, let us take the PD situation in Table 5.2, involving Coke and Pepsi. The reader may have wondered why we said earlier that the equilibrium in the PD situation was for both players to _defect ‘, with the result being that their combined payoff was less than the optimal combined payoff; both players end up worse than they could be if they co-operated ‘.

*Table 5.4. Iterated dominant strategy equilibrium*

| | | Pepsi: Maintain Price | Pepsi: Discount |
|---|---|---|---|
| Coke: Maintain Price | | 50 / 80 | 70 / –10 |
| Coke: Discount | | –10 / 70 | 10 / 10 |

Consider the situation from Coke ‘s viewpoint. If Pepsi maintains a high price, Coke is better off discounting, getting a payoff of 70 compared with 50. Similarly, if Pepsi discounts, Coke is still better off discounting, getting a payoff of 10 compared with -10. Thus, discounting is a dominant strategy for Coke. By the same line of reasoning, discounting is also the dominant strategy for Pepsi. We could also say that maintaining price is in this case a dominated strategy for both firms; this means that it will always give a lower or equal payoff, whatever the other player does.

Therefore, given the payoffs in Table 5.1, it is obvious that there is dominant strategy equilibrium, meaning that the strategies pursued by all players are dominant. In the situation in Table 5.2 both firms will discount, regardless of the fact that they will both be worse off than if they had maintained prices. By individually pursuing their self-interest each firm is imposing a cost on the other firm that they are not considering. It can therefore be said that in the PD situation the dominant strategy outcome is Pareto dominated. This means that there is some other outcome where at least one of the players is better off while no other player is worse off. However, Pareto domination considers total or social welfare; this is not relevant to the choice of strategy by each firm.

**b. Iterated dominant strategy equilibrium**

What would happen if one firm did not have a dominant strategy? This is illustrated in Table 5.4, which is similar to Table 5.2 but with one payoff changed. Coke now

gets a payoff of 80 instead of 50 if both firms maintain price. Although Pepsi 's dominant strategy is unchanged, Coke no longer has a dominant strategy. If Pepsi maintains price it is better off maintaining price, but if Pepsi discounts, Coke is better off also discounting. In this case, Coke can rule out Pepsi maintaining price (that is a dominated strategy), and conclude that Pepsi will discount; Coke can therefore iterate to a dominant strategy, which is to discount. Thus, the equilibrium is the same as before.

**c. Nash equilibrium**

The situation becomes more complicated when neither player has a dominant strategy. This is illustrated in Table 5.5. Note that this is no longer a Prisoner 's Dilemma, since the structure of the payoffs has changed.

Table 5.5. Game with no dominant strategy

| | | Pepsi | |
|---|---|---|---|
| | | Maintain Price | Discount |
| Coke | Maintain Price | 50<br>60 | 70<br>15 |
| | Discount | 10<br>75 | 5<br>10 |

In this situation, Coke is better off discounting if Pepsi maintains price, but is better off maintaining price if Pepsi discounts. The same is true for Pepsi. There is no single equilibrium here. Instead we have to use the concept of a Nash equilibrium. This represents an outcome where each player is pursuing their best strategy in response to the best-reply strategy of the other player. This is a more general concept of equilibrium than the two equilibrium concepts described earlier; while it includes dominant strategy equilibrium and iterated dominant strategy equilibrium, it also relates to situations where the first two concepts do not apply. There are two such equilibria in Table 5.5, as already seen:

1. If Coke maintains price, Pepsi will discount; and, given this best response, Coke 's best reply is to maintain price.
2. If Coke discounts, Pepsi will maintain price; and, given this best response, Coke 's best reply is to discount.

The same equilibria could also be expressed from Pepsi 's point of view:

1. If Pepsi discounts, Coke will maintain price; and, given this best response, Pepsi ‘s best reply is to discount.
2. If Pepsi maintains price, Coke will discount; and, given this best response, Pepsi ‘s best reply is to maintain price.

Coke will prefer the second strategy, while Pepsi will prefer the first. However, there is no further analysis that we can perform to see which of the two equilibria will prevail. This presents a problem for strategy selection if the game is repeated, as will be seen later. The concept of a Nash equilibrium is an extremely important one in game theory, since situations frequently arise where there is no dominant strategy equilibrium or iterated dominant strategy equilibrium.

## 5.4. Oligopoly models

There are various static models of oligopoly that have been proposed. The models that will be considered at this point are the Cournot model, the Bertrand model and the contestable markets model.

Although the Cournot and Bertrand models were developed independently, are based on different assumptions, and lead to different conclusions, there are some common features of the models that can be discussed at this stage before each is examined in detail. These features are easier to discuss in a two-firm framework, meaning a duopoly. Both models consider the situation where each firm considers the other firm ‘s strategy in determining its own demand function. The other firm ‘s strategy is considered to relate to either the output or price variable. Thus, these demand functions can be expressed as reaction or response curves which show one firm ‘s strategy, given the other firm ‘s strategy. The equilibrium point is where the two response curves intersect, meaning that the two firms ‘strategies coincide. To understand what all this is about we now have to consider each model separately.

### 5.4.1. The Cournot model

This model, originally developed in 1835, initially considered a market in which there were only two firms, A and B. In more general terms we can say that the Cournot model is based on the following assumptions:

- ✓ There are few firms in the market and many buyers.
- ✓ The firms produce homogeneous products; therefore, each firm has to charge the same market price (the model can be extended to cover differentiated products).

- ✓ Competition is in the form of output, meaning that each firm determines its level of output based on its estimate of the level of output of the other firm. Each firm believes that its own output strategy does not affect the strategy of its rival(s).
- ✓ Barriers to entry exist.
- ✓ Each firm aims to maximize profit, and assumes that the other firms do the same.

An essential difference between this situation and the ones considered until now is that strategies are continuous in the Cournot model. This allows a more mathematical approach to analysis. The situation can be illustrated by using the example relating to the cartel in the previous chapter. In that case the market demand was given by P = 400 - 2Q and each firm had constant marginal costs of £40 and no fixed costs. We saw that the monopoly price and output were £220 and 90 units, while price and output in perfect competition were £40 and 180 units. The analytical procedure can be viewed as involving the following steps.

Step 1. Transform the market demand into a demand function that relates to the outputs of each of the two firms. Thus, we have:

$$P = 400 - 2(Q_A + Q_B)$$

$$P = 400 - 2Q_A - 2Q_B \quad \text{.....} \textit{Equation 5.1}$$

Step 2. Derive the profit functions for each firm, which are functions of the outputs of both firms. Bearing in mind that there are no fixed costs and therefore marginal cost and average cost are equal, the profit function for firm A is as follows:

$$\prod_A = (400 - 2Q_A - 2Q_B)\, Q_A - 40Q_A = 400Q_A - 2Q^2_A - 2Q_BQ_A - 40Q_A$$

$$= 360QA - 2Q2A - 2QBQA \quad \text{.....} \textit{Equation 5.2}$$

Step 3. Derive the optimal output for firm A as a function of the output of firm B, by differentiating the profit function with respect to $Q_A$ and setting the partial derivative equal to zero:

$$\frac{\partial \prod_A}{\partial Q_A} = 360 - 4Q_A - 2Q_B = 0$$

$$4Q_A = 360 - 2Q_B$$

$$Q_A = 90 - 0.5Q_B \quad \text{.....} \textit{Equation 5.3}$$

Strictly speaking, the value of $Q_B$ in this equation is not known with certainty by firm A, but is an estimate. Equation 5.3 is known as the best-response function or

response curve of firm A. It shows how much firm A will put on the market for any amount that it estimates firm B will put on the market. The second and third steps above can then be repeated for firm B, to derive firm B's response curve. Because of the symmetry involved, it can be easily seen that the profit function for firm B is given by:

$$\prod_B = 360Q_B - 2Q^2_B - 2Q_BQ_A \text{ .....} \textit{Equation 5.4}$$

And the response curve for firm B is given by:

$$Q_B = 90 - 0.5Q_A \text{ .....} \textit{Equation 5.5}$$

This shows how much firm B will put on the market for any amount that it estimates firm A will put on the market. The situation can be represented graphically, as shown in Figure 5.1.

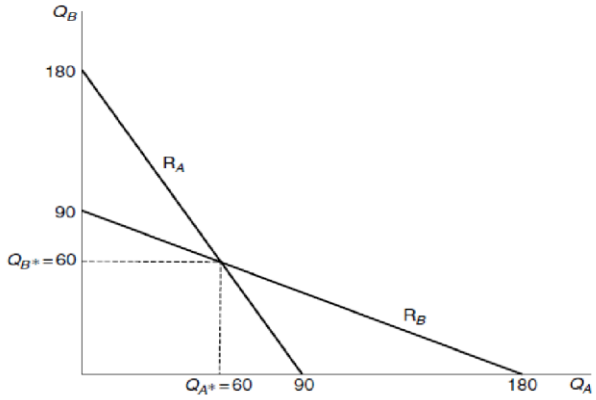


*Figure 5.1. Cournot response curves*

Step 4. Solve the equations for the best-response functions simultaneously to derive the Cournot equilibrium. The properties of this equilibrium will be discussed shortly.

$Q_A = 90 - 0.5Q_B$

$Q_B = 90 - 0.5Q_A$

$Q_A = 90 - 0.5 (90 - 0.5Q_A)$

$Q_A = 90 - 45 + 0.25Q_A$

$0.75Q_A = 45$

$Q_A = 60$

$Q_B = 90 - 0.5 (60)$

= 60 The market price can now be determined:

$P = 400 - 2(60+ 60)$

$P = £160$

We can now compare this situation with perfect competition, monopoly and cartels. This is shown in Table 5.6. A Cournot duopoly clearly does not make as much profit as a cartel involving the two producers; by colluding, the two firms could restrict output to increase profit. The reason for this is that, when one firm increases output, the price for the market as a whole is reduced.

*Table 5.6. Comparison of perfect competition, monopoly and Cournot dupoly*

| Market structure | Price (£) | Output in the industry | Profit in the industry (£) |
|---|---|---|---|
| Perfect competition | 40 | 180 | 0 |
| Monopoly (or cartel) | 220 | 90 | 16,200 |
| Cournot duopoly | 160 | 120 | 14,400 |

However, the firm does not consider the effect of the reduced revenue of the other firm, called the revenue destruction effect, since it is only concerned with maximizing its own profit. Thus, it expands its output more aggressively than a cartel would, since a cartel is concerned with the profit of the industry as a whole. This Cournot equilibrium is also called the Cournot–Nash equilibrium (CNE), since it satisfies the conditions stated in the last subsection regarding the nature of a Nash equilibrium. The CNE represents the situation where the strategies of the two firms _match ‘, and there will be no tendency for the firms to change their outputs; at any other pair of outputs there will be a tendency for the firms to change them, since the other firm is not producing what they estimated. The Cournot equilibrium is therefore a comparative static equilibrium.

This can be illustrated in the following example of the adjustment process. Let us assume that both firms start by producing 80 units. Neither firm will be happy with their output, since they are producing more than they want, given the other firm ‘s actual output. Say that firm A is the first to adjust its output. It will now produce 90-0.5 (80) = 50 units. Firm B will react to this by producing 65 units; firm A will then produce 57.5 units; firm B will then produce 61.25 units and so on, with the outputs converging on 60 units for each firm.

It should be noted that the Cournot–Nash equilibrium has stronger properties than other Nash equilibria. If strictly dominated strategies are eliminated, as shown in the above example, only one strategy profile remains for rational players, the

Cournot–Nash equilibrium. It can therefore be concluded that this represents a unique iterated strictly dominant strategy equilibrium of the Cournot game.

All the preceding analysis has been based on a two-firm industry. As the number of firms in the industry increases, the market price is reduced and market output increases. The reason for this is related to the revenue destruction effect described earlier. With more firms in the industry, any increase in output by one firm has a smaller effect on the market price, and on its own profit, but the effect on the combined revenues of all the other firms increase in comparison to the effect on its own profit.

### 5.4.2. The Bertrand models

This model dates back to 1883. The assumptions involved in the model are as follows:

1. There are few firms in the market and many buyers.
2. The firms produce homogeneous or differentiated products; therefore, each firm has to charge the same market price in the case of homogeneous products, but there is some scope for charging different prices for differentiated products.
3. Competition is in the form of price, meaning that each firm determines its level of price based on its estimate of the level of price of the other firm. Each firm believes that its own pricing strategy does not affect the strategy of its rival(s).
4. Barriers to entry exist.
5. Each firm has sufficient capacity to supply the whole market.
6. Each firm aims to maximize profit, and assumes that the other firms do the same. It can be seen that the first, fourth and fifth assumptions are the same as for the Cournot model but that competition is in the form of price rather than output. We can begin by taking the Cournot equilibrium analysed above, where each firm charge £160 and sells 60 units. If a firm in Bertrand competition believes that its rival will charge £160 it can undercut its rival slightly, by charging say £159, and capture the whole market, since we have assumed that the product is homogeneous, and that firms have sufficient capacity to do this. This action would considerably increase its profit. However, any rival will then react by undercutting price again, to say £158, and again capture the whole market.

Thus, no matter how many or how few firms there are in the market, price reductions will continue until the price has been forced down to the level of marginal cost. The conclusion, therefore, is that, with homogeneous products, the Bertrand equilibrium is identical with that in perfect competition, with no firm making any economic or supernormal profit. The more important situation in practical terms, where the products are differentiated, is discussed shortly.

Why is there such a big difference between the Cournot and Bertrand models in terms of their conclusions regarding prices, outputs and profits?

The assumptions underlying the models are fundamentally different. In the Cournot model, firms make production decisions that tend to involve long-run capacity decisions; once these decisions have been made, firms tend to sell their output at whatever price it will fetch, thus avoiding price competition. In the Bertrand model, price competition is much more intense, and each firm believes that it can steal a rival ‘s market by cutting its price just slightly. Production is very flexible under these assumptions, so that a firm can increase output considerably in order to supply a large number of new customers. When products are differentiated, as they are in most oligopolistic markets, the analysis of Bertrand competition becomes more complex, and resembles the Cournot model in some respects. First of all, a firm will not lose its whole market if a rival undercuts its price, it will lose only some of its customers, depending on the cross-elasticity of demand. Assuming a two-firm situation again for simplicity, the model is based on each firm having a demand function related both to its own price and to that of its competitor. This corresponds to the Cournot situation where each firm has a demand function related both to its own price and to the output of the competitor. Again similarly to the Cournot analysis, profit functions for each firm are derived (assuming that each firm‘s cost function is known) and these functions are maximized by differentiating them and setting the derivatives equal to zero. The resulting equations yield the best-response functions of the firms, which can then be solved simultaneously to derive the equilibrium prices.

Thus, the analysis is essentially similar to the four-step procedure described in the Cournot situation. It is useful at this stage to illustrate the procedure by using an example.

**Step 1.** Let us take two firms, A and B, with the following demand functions:

$Q_A$ - 60 - $4P_A$ + $2.5P_B$ $Q_B$ - 50 - $5P_B$ + $2P_A$ where Q is in units and P is in Birr. In this case the individual demand functions are given; market demand would be the sum of the two functions. Firm A has marginal costs of Birr 5 and Firm B has marginal costs of Birr 4. Note that these marginal costs can be different because the product is differentiated, unlike the Cournot case. It is assumed for simplicity that there are no fixed costs.

**Step 2**. Derive profit functions for each firm (Profit = (P-AC) Q):

$\Pi_A = (P_A-5)\ (60-4P_A+2.5P_B)$

$= 60P_A-4P2_A+2.5P_AP_B-300+20P_A-12.5P_A$

$= 80P_A-4P2_A+2.5P_AP_B-300-12.P_A$

$\Pi_B = (P_B-4)\ (50-5P_B+2P_A)$

$=50P_B-5P^2_B+2P_BP_A-200+20P_B-8P_A$

$=70P_B-5P^2_B+2P_BP_A-200+8P_A$

**Step 3.** Differentiate the profit functions and set equal to zero in order to derive best-response functions:

$\frac{\partial \Pi_A}{\partial \Pi A} = 80-9P_A+2.5P_B-0$

$P_A=10+0.3125P_B$

$\frac{\partial \Pi_B}{\partial \Pi_B} = 70-10P_B+2P_A-0$

PB =7+0.2PA

**Step 4.** Solve best-response functions simultaneously to derive the equilibrium price:

$P_A$ =10 + 0.3125(7 + 0.2) PA

$P_A = 12.1875 + 0.0625P_A$

$0.9375P_A = 12.1875$

$P_A$ =*Birr*13

$P_B = 7 +0.2(13)$

$P_B$ =*Birr* 9.60

Note that the response curves are positive sloping in this case. This means that the higher the price that one firm thinks the other firm will charge, the higher the price it will itself charge. Some readers might think at this point that the kind of analysis above is too theoretical and abstract to be of much use in practice. This is not true. In an empirical study of the cola market in the United States from 1968 to 1986, precisely the methodology above was used to estimate the prices that Coca-Cola and Pepsi should charge. It was concluded that Coca-

Cola ‘s price should be $12.72 and Pepsi ‘s price should be $8.11. When one compares these estimates with the actual average prices over the period of $12.96 and $8.16, one can see that the Bertrand model is capable of forecasting actual pricing behavior with remarkable accuracy.

## 5.5. Dynamic games

Many business scenarios tend to involve sequential moves rather than simultaneous moves. An example is the decision to invest in new plant. Sometimes firms can change a simultaneous game into a sequential game, whereas in other situations the game is naturally of a sequential type, for example when a management makes a wage offer to a labor union. Consider a situation of the first type, where there are two firms competing in an oligopolistic industry. Both firms are considering an expansion of their capacity in order to increase market share and profit. The resulting profits (in millions of pounds) are shown in Table 5.7. It is assumed for simplicity that only two strategies are available to each firm, whereas, in reality, different scales of expansion would probably be possible; thus, we are making the game into one of discrete rather than continuous strategies. This simplification does not affect the nature of the situation.

Equilibrium

In a simultaneous game Firm, A has the dominant strategy of making no change and Firm B iterates to the strategy of expanding. Thus, Firm A will get a payoff of 70 and B will get its maximum payoff of 40. Firm A is not happy with this situation since it is better off if Firm B makes no change, but how can it persuade Firm B to do this? By moving first and taking the strategy of expanding. This transforms the game into a sequential one. Dynamic games are best examined by drawing a game tree. The relevant game tree for Table 5.7 is shown in Figure 5.2. Such a game tree represents the game in extensive form. An extensive-form game not only specifies the players, possible strategies, and payoffs, as in the normal-form game, but also specifies when players can move, and what information they have at the time of each move.

In order to analyze this game tree, we must derive the sub game perfect Nash equilibrium (SPNE). This is the situation where each player selects an optimal action at each stage of the game that it might reach, believing the other player(s)

will act in the same way. Decision nodes for each firm are shown by rectangles, and payoffs are shown with Firm A first in each pair.

*Table 5.7. Transforming a simultaneous game into a sequential game*

| | | Firm B: Expand | Firm B: No Change |
|---|---|---|---|
| Firm A: Expand | | 20 / 50 | 25 / 85 |
| Firm A: No Change | | 40 / 70 | 30 / 95 |

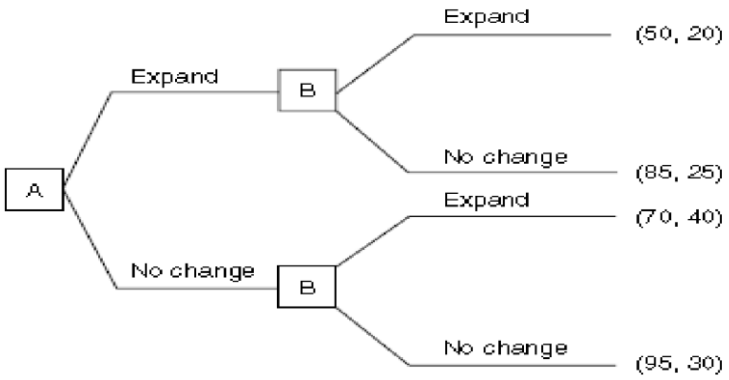


*Figure 5.2. Game tree for capacity expansion game*

The SPNE is obtained by using the _fold-back ‘or backwards induction method. This means that we proceed by working backwards from the end of the tree, at each stage finding the optimal decision for the firm at that decision node. Thus, if Firm A expands, the best decision for Firm B is to make no change (payoff of 25, compared with 20 for expanding). If Firm A makes no change the best decision for Firm B is to expand, with a payoff of 40 compared with 30 for making no change. Knowing this, Firm A can now make its original decision. If it expands, Firm B will make no change, and Firm A will get a payoff of 85; if it makes no change, Firm B will expand and Firm A will get a payoff of 70. Therefore, Firm A, acting first, will make the decision to expand. This yields an improvement of 15 compared with the original payoff of 70. Thus the apparently strange strategy of expansion by Firm A makes sense.

The game tree shown in Figure 5.2 applies to the situation where strategies are discrete. If strategies are continuous, game trees are less useful, since in this case

strategies are generally represented as best-response functions, as we have already seen in the Cournot and Bertrand cases. However, game trees can still be an aid to clarifying the order of actions taken, and therefore the order of analysis.

**Strategic moves and commitment**

The move by Firm A is an example of a strategic move. This is an action _designed to alter the beliefs and actions of others in a direction favorable to yourself '. The decision by Firm A to expand may seem strange, for two reasons:

1. Firm A 's profit from making no change is greater than its profit from expanding, regardless of what strategy Firm B takes.
2. Moving first and investing in expansion limits Firm A's options and creates inflexibility.

This limitation of one 's own options, combined with the credible commitment to do so, is the key feature of strategic moves. It has been called the paradox of power. Let us consider a few realistic examples from various walks of life. Examples of making commitments are: protesters handcuffing themselves to railings, and _doomsday 'nuclear devices designed to automatically strike the enemy if attacked. An archetypal example of commitment is the burning of one 's boats, as when Cortes landed in Mexico to conquer the Aztec empire. This caused his soldiers to fight harder than otherwise, since they knew that they had no alternative.

The situation described in Figure 5.2 is an example of commitment. It illustrates the important point that strategies that may initially appear strange can in fact make sense. The inflexibility factor can cause other players to behave in ways in which they otherwise would not, and which favour the player making the commitment. For a commitment to be successful in causing other players to change their behavior it must be visible, understandable and credible. The first two characteristics are fairly self-explanatory, but the aspect of credibility needs to be examined in some detail. For a strategic move to be credible it must be more in the player 's interest to carry out the move than not to carry it out. A key factor here is irreversibility. Consider the situation in Table 5.7. If Firm A merely announces in the media its intention to expand, this action has little credibility because it is easy to reverse. However, if Firm A actually starts building operations, then this action is much costlier to abandon. This is particularly true if the investment activity involves considerable sunk costs. Thus, although it was stated that sunk costs are not

important in decision-making, in this context sunk costs are an important factor. They create the kind of inflexibility that is necessary for a commitment to be credible. There are a number of factors that can create such irreversibility.

1. Writing contracts. These legal agreements can make it difficult and costly for a player to change its actions later. Thus, although we have seen that contracts are important in order to prevent the other party from opportunistic behavior, by limiting their actions, they can also be useful in making credible

2. Reputation effects. The threat of losing one ‘s reputation can serve as a credible commitment. For example, when one firm offers to provide a service to another firm, and moral hazard is involved because the performance of the service cannot be immediately observed, the firm offering the service can claim that it cannot afford to provide a bad service because it would lose its reputation.

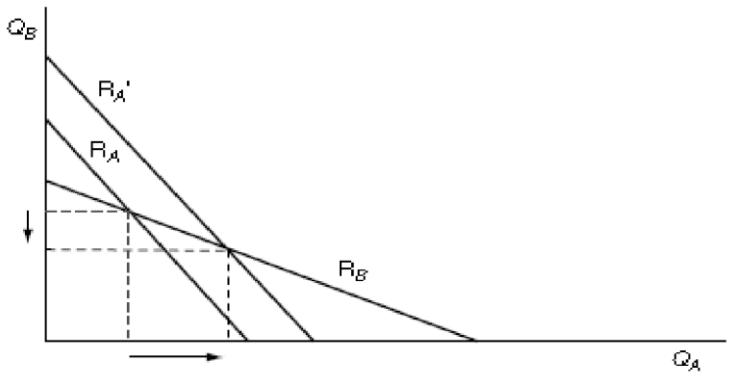


*Figure 5.3. The effect of commitment on Cournot response curves*

3. Burning one ‘s bridges. The example of Cortes in Mexico shows the effect of this type of action. However, the action does not have to be as drastic as this. For a politician, a publicized policy statement may be sufficient, although politicians have been known to do _U-turns ‘.

4. Breaking off communications. This again prevents the player from changing their actions at a later date. However, such a move can also create other problems in terms of discovering the effectiveness of one ‘s strategy in relation to the other player(s). a. Cournot commitment

The kind of commitment related to the situation described in Table 5.7 can also be illustrated using the analysis of Cournot competition. This is competition in terms of output, and a credible commitment to expand will shift Firm A ‘s response curve to the right. For any given output of Firm B, Firm A will now produce more than before, as shown in Figure 5.3. This in turn causes a change in the equilibrium

situation, with Firm A producing more output and Firmer B producing less output than before the commitment, just as we saw using the game tree analysis.

b. Bertrand commitment

The commitment to expand is only one type of commitment. Let us return to the pricing competition in Figure 5.3. This was originally a simultaneous game involving a Prisoner ‘s Dilemma; the dominant strategy for Pepsi (Firm B) was to discount, thus leading Coke also to discount. The situation is reproduced in general terms in the game matrix of Table 5.8.

*Table 5.8. Prisoner’s Dilemma in price competition*

| | | Firm B: Maintain Price | Firm B: Discount |
|---|---|---|---|
| Firm A | Maintain Price | 50 / 80 | 70 / –10 |
| | Discount | –10 / 70 | 10 / 10 |

As we have already seen, both players end up worse off in this situation than they would be if they both maintained prices. How can a firm credibly commit to a strategy of maintaining price? Merely announcing an intention to maintain price is not sufficient, since this has no credibility, given the dominant strategies of both firms. An ingenious solution to this problem has been commonly implemented by many firms in oligopolistic situations. This involves a firm using a _most favoured customer clause ‘(MFCC). Essentially what this involves is a guarantee to customers that the firm will not charge a lower price to other customers for some period in the future; if it does, it will pay a rebate to existing customers for the amount of the price reduction, or sometimes double this amount. This is particularly important in consumer durable markets. The reason why this strategy is ingenious is that it serves a dual purpose.

1. Ostensibly, it creates good customer relations. Many customers are concerned when they are considering buying a new consumer durable that the firm will reduce the price of the product later on. This applies particularly when there are rapid changes in technology and products are phased out over relatively short periods, for example computers and other electronics products.

2. The MFCC creates a price commitment. It would be expensive for the firm to reduce price at a later stage, since it would have to pay rebates to all its previous customers. Thus, other firms are convinced that the firm will maintain its price, and this causes prices to be higher than they would be without such commitment, as is seen below.

A commitment by Firm A to maintain price in this situation will not achieve the desired effect; Firm B will discount, and Firm A will end up worse than before. However, there is scope for Firm B to make such a commitment; in this case it would be in Firm A 's best interest also to maintain price, obtaining a payoff of 80, compared with 70 if it discounts. Thus, Firm B will end up better off than originally without such a commitment, obtaining a payoff of 50, compared with 10. Ironically, contrary to what customers might have expected, the MFCC by Firm B causes prices to be higher than they otherwise would have been.

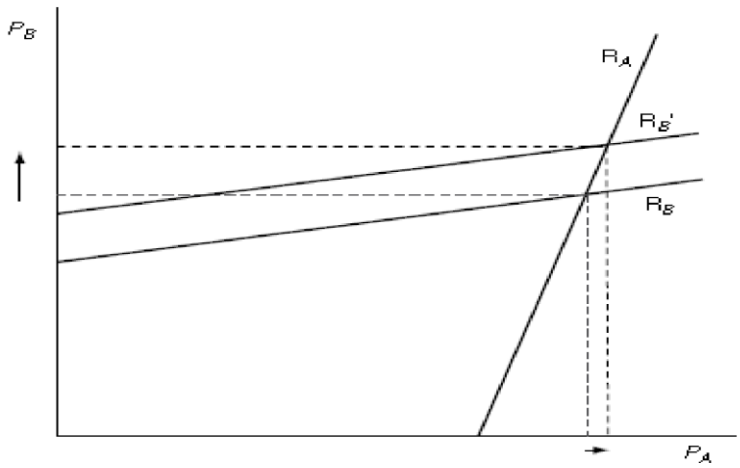


*Figure 5.4. The effect of commitment on Bertrand response curves*

The situation here can again be represented graphically, this time in terms of Bertrand competition. The commitment by Firm B to maintain price has the effect of shifting its reaction curve upwards, meaning that for any given price charged by Firm A, it will charge a higher price than before. The effect is to change the equilibrium, with both firms charging a higher price than before the commitment. This is illustrated in Figure 5.4. We have now considered two kinds of commitment, one in terms of output and one in terms of price. Another type of output commitment is the investment in excess capacity, which serves as a deterrent to entry. Potential entrants recognize the ability of an existing firm to increase output substantially, at low marginal cost; thus, driving down the market price. The incumbent firm can

afford to do this because of its low marginal cost; thus, the threat to potential entrants is credible. Other types of commitment can also be considered, as well as other methods of achieving credibility. For example, commitments can also be made in terms of product quality and advertising expenditure

---

***Self Assessment Exercise***

Explain the differences between the Cournot and Bertrand models of competition; why are these models not true models of interdependent behaviour?

1. Explain the following terms:
   a. Dominant strategy
   b. Nash equilibrium
   c. Most favoured customer clause
   d. Mixed strategies.
2. Explain the relationship between strategic moves, commitment and credibility.
3. Explain how you would formulate a strategy for playing the paper–rock– scissors game on a repeated basis.
4. Explain why it makes a difference in a repeated game if the end of the game can be foreseen.

# CHAPTER SIX
# 6.MARKETFAILUREANDGOVERNMENT INTERVENTION

***Objectives***

- *To explain why government policy is important for managerial decision making.*
- *To explain the concept of market failure, transaction costs and externalities and their implications for both governments and businesses.*
- *To explain the concept of public goods and their relationship with externalities.*

## 6.1. Introduction

Throughout most of this material, we have treated the market as a place where firms and consumers come together to trade goods and services with no intervention from government. But as you are aware, rules and regulations that are passed and enforced by government enter into almost every decision firms and consumers make. As a manager, it is important to understand the regulations passed by government, why such regulations have been passed, and how they affect optimal managerial decisions.

We will begin by examining four reasons why free markets may fail to provide the socially efficient quantities of goods: (1) market power, (2) externalities, (3) public goods, and (4) incomplete information. Our analysis includes an overview of government policies designed to alleviate these ―market failures‖ and an explanation of how the policies affect managerial decisions. The power of politicians to institute policies that affect the allocation of resources in markets provides those adversely affected with an incentive to engage in lobbying activities. We will illustrate the underlying reasons for these types of rent-seeking activities. Finally, we will examine how these activities can lead politicians to impose restrictions such as quotas and tariffs in markets affected by international trade.

## 6.2. Market failure

One of the main reasons for government involvement in the marketplace is that free markets do not always result in the socially efficient quantities of goods at socially efficient prices. In this section, we will consider why markets do not always lead to socially efficient outcomes and examine how government policies designed to correct ―market failures‖ affect managerial decisions. We will begin by examining market failure due to the presence of market power.

## 6.2.1. Market Power

Social welfare—defined as the sum of consumer and producer surplus—is maximized at the output where price equals marginal cost. The socially efficient price and output arises naturally if the industry is perfectly competitive. In contrast, a firm that has *market power* produces less than the socially efficient level of output because it charges a price that exceeds its marginal cost of production. In such instances, the value to society of another unit of the good is greater than the cost of the resources needed to produce that unit; there would be a net gain to society if additional output were produced. In these instances, government may intervene in the market and regulate the actions of firms in an attempt to increase *social welfare.*

To see the potential benefits of government intervention in a market, consider a market serviced by a monopoly. Figure 6.1 shows the monopolist ‘s demand, marginal cost, and marginal revenue curves. Assuming the monopolist must charge the same price to all consumers in the market, the profit-maximizing output is $Q^M$ units, and these units are sold at the monopoly price of $P^M$. At this price, consumers pay more for the last unit of output than it costs the producer

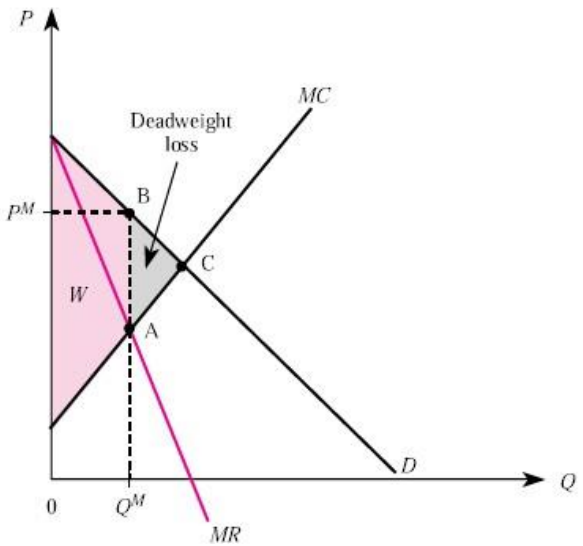


*Figure 6.1. Welfare and Deadweight Loss under Monopoly*

to manufacture and sell it. Total social welfare under monopoly is the region labeled *W* in Figure 6.1.

Notice in Figure 6.1 that the area of triangle ABC is the *deadweight loss* of the monopoly, welfare that would have accrued to society if the industry were perfectly

competitive but is not realized because of the market power the monopolist enjoys. The failure of the market to fully maximize social welfare is due to market power; the deadweight loss triangle provides a measure of this welfare loss to society.

 **Antitrust policy**

The government uses antitrust policy to enact and enforce laws that restrict the formation of monopolies. The rationale for these policies is that by preventing monopoly power from emerging, the deadweight loss of monopoly can be avoided. In some instances, however, the presence of economies of scale makes it desirable to allow the formation of a monopoly. In these cases, government attempts to reduce the deadweight loss by regulating the price charged by the monopolist.

 Price Regulation

In the presence of large economies of scale (as is the case for some utility companies), it may be desirable for a single firm to service a market. In these instances, government may allow a firm to exist as a monopoly but choose to regulate its price to reduce the deadweight loss. In this section, we will see how such regulation affects managerial decisions and social welfare.

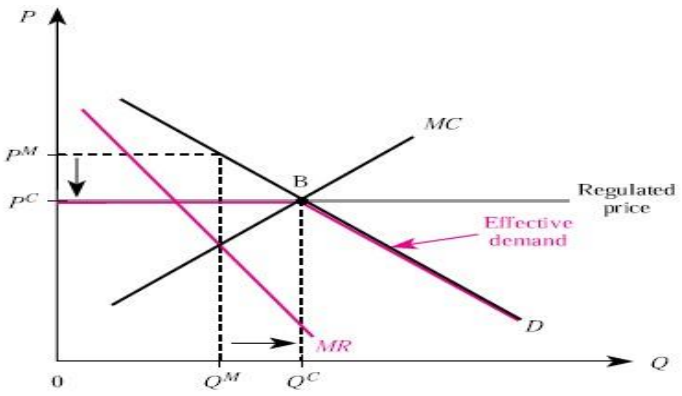


*Figure 6.2. Regulating a Monopolist's Price at the Socially Efficient Level*

Consider the situation depicted in Figure 6.2, where an unregulated monopolist produces $Q^M$ units of output at a price of $P^M$. A competitive industry would produce $Q^C$ units, where marginal cost intersects the demand curve. Suppose the government imposed and enforced a regulated price of $P^C$, which corresponds to the price a

competitive industry would charge for the product given identical demand and cost conditions. How should the manager respond to maximize the firm ‘s profits?

The monopolist cannot legally charge a price above $P^C$, so the maximum price it can charge for units less than $Q^C$ is $P^C$. For units above $Q^C$, the maximum price it can charge is the price along the inverse demand curve, since the amount consumers are willing to pay is less than the ceiling. As a consequence, the effective inverse demand curve of the monopolist is given by $P^CBD$. Notice that for points to the left of B, the demand curve is horizontal, just as it is for a perfectly competitive firm. But if the monopolist wishes to sell more than $Q^C$ units of output, it can do so only by lowering price below $P^C$. Since the monopolist can sell each unit up to $Q^C$ at a price of $P^C$, the marginal revenue for these units is simply $P^C$: Each additional unit of output up to $Q^C$ adds exactly $P^C$ to the firm ‘s revenue. In effect, the ceiling creates a situation where the demand curves the monopolist faces is just like that of a perfectly competitive firm for these output levels. To maximize profits, the regulated monopolist will produce where the marginal revenue of the effective demand curve ($P^C$) equals marginal cost, which in this case is at point B. This corresponds to an output of $Q^C$. Thus, when the monopolist ‘s price is regulated at $P^C$ in Figure 6.2, the firm maximizes profits by producing $Q^C$ units and selling them at the regulated price of *PC.*

### 6.2.2. Externalities

Unfortunately, some production processes create costs for people who are not part of the production or consumption process for the good. These external costs are called *negative externalities.* The most common example of a negative externality is pollution. When a firm creates wastes that either do not easily biodegrade or have harmful effects on other resources, it does not pay the full cost of production. For example, a firm that produces textiles usually creates waste products that contain dioxin, a cancer-causing chemical. When a textile manufacturer can dispose of this waste ―for free‖ by dumping it into a nearby river, it has an incentive to dump more waste into the river than is socially optimal. While the firm benefits from dumping waste into the river, the waste reduces the oxygen content of the water, clogs normal waterway routes, and creates reproduction problems for birds, fish, reptiles, and

aquatic animals. These results negatively affect people who are not involved in the production or consumption process.

To see why the market fails to provide the efficient level of output when there are externalities in production, consider Figure 6.3. If a firm emits pollutants into the water as a by-product of producing steel, a cost, or negative externality, is borne by members of society.

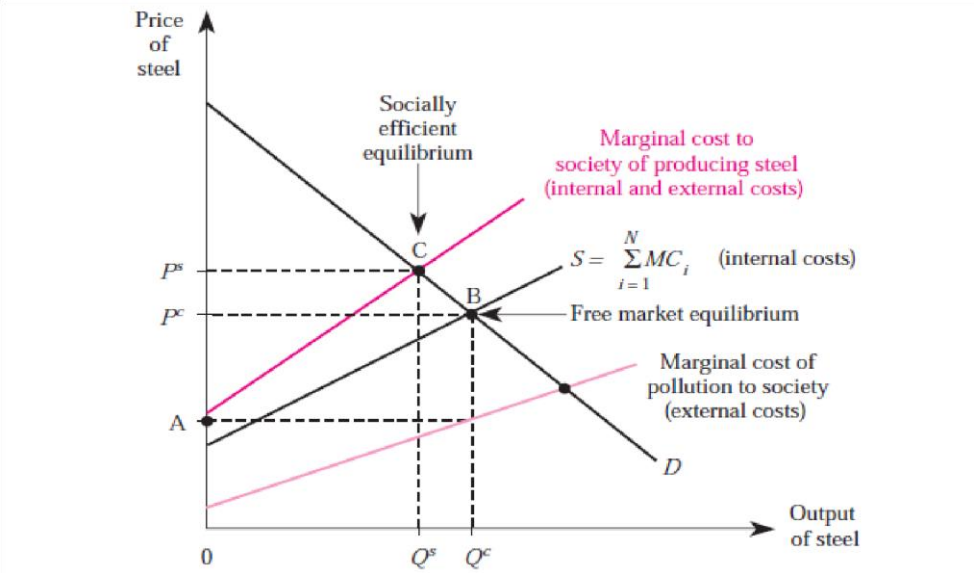


*Figure 6.3. The socially efficient equilibrium in the presence of external cost*

Figure 6.3 shows the negative externality as the marginal cost of pollution to society. This cost represents the cost to society of dirtier water due to increases in steel production. The production of only a little steel results in only minor damage to water, but as increasing amounts of steel are produced, more and more pollutants collect in the water. The marginal cost of pollution to society thus increases as more steel is produced.

Assuming the market for steel is perfectly competitive, the market supply curve is *S* in Figure 6.3, which is the sum of the marginal costs of the firms producing in the industry. The supply curve is based on the costs paid by the steel firms; thus, if they are allowed to dump pollutants into the water for free, the market equilibrium is at point B, where the market demand and supply curves intersect. The result is that $Q^C$ units will be produced and purchased at a price of $P^C$ per unit of steel.

However, at this quantity of output, society pays a marginal price of *A,* on top of the price of $P^C$ paid to the steel firms. This amount is the additional cost to society of the pollution. In particular, since the firms are dumping pollutants into the water for free, the cost of pollution is not internalized by those who buy and sell steel; rather, society pays for the dumping of pollutants by having to endure polluted water. If the firms were to consider the cost of pollution to society, the sum of their marginal cost curves would be the vertical sum of the supply curve and the marginal cost of pollution to society. This sum is shown as the marginal cost to society of producing steel in Figure 6.3. The socially efficient level of output, which considers all the costs and benefits of producing steel, is at point C, where the marginal cost to society of producing steel intersects the market demand curve. The socially efficient level of steel output is $Q^S$, which is less than the output produced in the perfectly competitive market. The socially efficient price of steel is $P^S$, which is greater than the perfectly competitive price of $P^C$. In other words, in the presence of external costs, the market equilibrium output is greater than the socially efficient level, and the market price is below the socially efficient level. In effect, consumers get to purchase too much output at too low a price.

The basic reason for the ―market failure‖ is the absence of well-defined *property rights;* the steel firms believe they have the right to use the river to dump waste, and environmentalists believe they have the right to a clean river. This failure often can be solved when government defines itself to be the owner of the environment. It can then use its power to induce the socially efficient levels of output and pollution. To induce the socially efficient level of output, government may force firms to internalize the costs of emitting pollutants by enacting policies that shift the internal cost of production up to where it actually equals the social cost of production.

### 6.2.3. Public Goods

Another source of market failure is the provision of *public goods,* goods that are non-rival and non-exclusionary in nature and therefore benefit persons other than those who buy the goods. Public goods differ from most goods you consume, which are rivalrous in nature. This simply means that when you consume the good, another person is unable to consume it as well. For example, when you buy and wear a pair of shoes to protect your feet, you prevent someone else from wearing the same pair;

the consumption of shoes is rivalrous in nature. *Non-rival* goods include radio signals, lighthouses, national defence, and protecting the environment. When you receive a radio signal in your car, you do not prevent other drivers from picking up the same station in their cars. This is in sharp contrast to your purchasing a pair of shoes. The second aspect of a public good is that it is *non-exclusionary:* Once a public good is made available, everyone gets it; no one can be excluded from enjoying the good. Most goods and services are by nature exclusionary. For example, when a car manufacturer produces a car, it can keep people from using the car by putting a lock on the door and giving the key only to the person who is willing to pay for the car. Goods and services such as clean air, national defence, and radio waves are non-exclusionary goods. For example, when the air is clean, everyone gets to consume the clean air; it cannot be allocated to a single person.

What is it about public goods that lead the market to provide them in inefficient quantities? The answer is that since everyone gets to consume a public good once it is available, individuals have little incentive to purchase the good; rather, they prefer to let other people pay for it. Once it becomes available, they can ―free ride‖ on the efforts of others to provide the good. But if everyone thinks this way, no one will buy the good and it will not be available. One person alone may be unable to afford to purchase the good.

### 6.2.4. Incomplete Information

For markets to function efficiently, participants must have reasonably good information about things such as prices, quality, available technologies, and the risks associated with working in certain jobs or consuming certain products. When participants in the market have *incomplete information* about such things, the result will be inefficiencies in input usage and in firms ‘output.

Consider the consumption of cigarettes. If individuals are not told that cigarettes are hazardous, some people who currently do not smoke because of the known health risks would smoke out of ignorance of the dangers of smoking. The decision to smoke would be based on incomplete information about the dangers of smoking. For reasons such as these, government serves as a provider of information in many markets, dispensing information to consumers about the ingredients of certain foods,

the dangers of certain products and drugs, and the like. Firms print some of this information on the labels of their products due to regulations imposed by government. Government even regulates the work environment by ensuring that workers are aware of the dangers of chemicals such as asbestos and the benefits of precautions such as wearing hard hats in construction jobs. In these instances, the regulations are carried out by the Occupational Safety and Health Administration of nations.

One of the more severe causes of market failure is asymmetric information, a situation where some market participants have better information than others. The presence of asymmetric information can lead buyers to refuse to purchase from sellers out of fear that the seller is willing to get rid of the product because it is worth less than they are willing to pay. In the extreme case, the market can collapse altogether. For this reason, several government policies are designed to alleviate the problems caused by asymmetric information. A few of the policies that affect managers are discussed next.

- *Rules against Insider Trading*

One example of a government regulation designed to alleviate market failures due to asymmetric information is the law against *insider trading* in the stock market. The purpose of the law is to ensure that asymmetric information (better information by insiders) does not destroy the market by inducing outsiders to stay out of it.

- *Certification*

Another policy government uses to disseminate information and reduce asymmetric information is the certification of skills and/or authenticity. The purpose of *certification* is to centralize the cost of gathering information. All licensing done by the government falls under certification; this includes all non-profit organizations, such as charities. Certification can also be a set of minimum standards, such as those for schools and physicians. The purpose is to assure consumers that the products or services have been certified as meeting a certain set of minimum standards. Without a central authority to fulfil this information-gathering role, each individual would have to pay the cost of gaining knowledge about the quality of a product or service.

 *Truth in Advertising*

Often firms have better information about their products than do consumers. This advantage may give firms an incentive to make false claims about the merits of their products to capitalize on consumers 'lack of information. In some instances, such practices can lead consumers to switch from one firm 's product to a competitor 's product. In extreme cases, the asymmetric information can induce consumers to ignore advertising messages altogether, out of fear that the messages are false. Government often can alleviate these market failures by regulating the advertising practices of firms. Advertising regulation, which encourages *truth in advertising*, usually is enforced by civil suits.

 *Enforcing Contracts*

Another way government solves the problems of asymmetric information is through *contract enforcement.* Contracts are written to keep the parties from behaving opportunistically in the final period of a game. For example, suppose your boss ―promised‖ your payment for labour services at the end of the month. After you have worked for a month, your boss refuses to pay you in effect gaining a month 's worth of your labour for free. These types of problems do not arise when reputation is important or when there is the potential for repeated interaction among the parties. In these instances, the onetime gain to behaving opportunistically will be more than offset by future losses. One solution to this problem is for government to enforce contracts. By enforcing contracts, government effectively solves the ―end-of-period‖ problem by requiring dishonest people to honour the terms of contracts.

## 6.3. Rent seeking

The preceding analysis shows how government policies can improve the allocation of resources in the economy by alleviating the problems associated with market power, externalities, public goods, and incomplete information. It is important to note, however, that government policies generally benefit some parties at the expense of others. For this reason, lobbyists spend considerable sums in attempts to influence government policies. This process is known as *rent seeking.*

To illustrate rent seeking and its consequences, suppose a politician has the power to regulate the monopoly in Figure 6.4. The monopoly currently charges a price of $P^M$, produces $Q^M$ units of output, and earns profits described by the shaded region A. At the monopoly price and output, consumer surplus is given by triangle C.

If consumers could persuade the politician to regulate the monopolist 's price at the competitive level ($P^C$), the result would be an output of $Q^C$. If this happened, the monopoly would lose all of its profits (rectangle A). Consumers, on the other hand, would end up with total consumer surplus of regions A, B, and C.

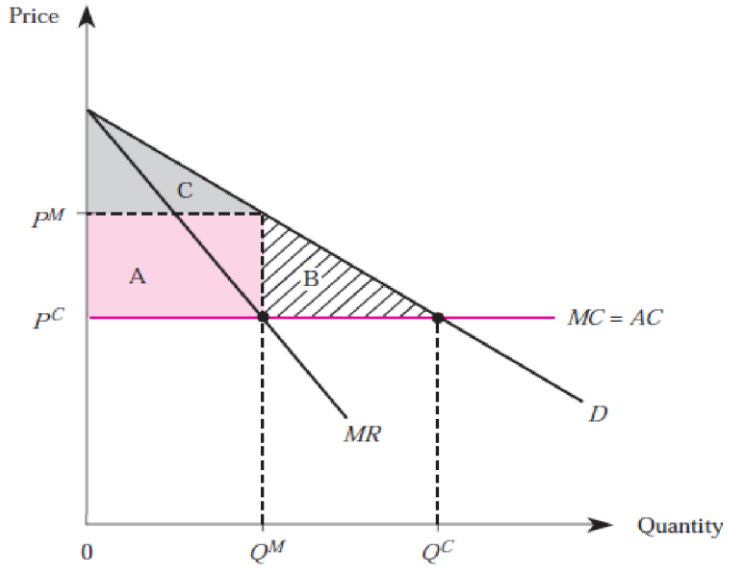


*Figure 6.4. The incentive to engage in rent-seeking activities*

Since the monopolist stands to lose rectangle A if the regulation is imposed, it has an incentive to lobby very hard to prevent the regulation from being imposed. In fact, the monopolist is willing to spend up to the amount A to avoid the regulation. These expenditures may be in the form of legal activities, such as campaign contributions or wining and dining the politician, or illegal activities such as bribes. Notice that the consumers in Figure 6.4 also would be willing to spend money to persuade the politician to regulate the monopoly. In fact, as a group they would be willing to spend up to A + B to impose the competitive solution, since this is the additional consumer surplus enjoyed when the price is $P^C$. Of course, each

individual consumer stands to gain much less than the group (regulation is a public good in that it benefits all consumers).

Consequently, each consumer has an incentive to ―free ride, and in the end the amount the consumers spend as a group will be very low. The monopolist, on the other hand, is a single entity. Avoiding regulation is not a public good to the monopolist; the monopolist will receive private gains if it can avoid the legislation. As a result, the monopolist generally will spend much more on lobbying activities than the consumers and thus will often avoid legislation by engaging in rent-seeking activities.

## 6.4. Government Policy and International Markets

Sometimes rent seeking manifests itself in the form of government involvement in international markets. Such policies usually take the form of tariffs or quotas that are designed to benefit specific firms and workers at the expense of others. In this section, we will examine how government tariff and quota policies affect managerial decisions.

### 6.4.1. Quota

The purpose of a *quota* is to limit the number of units of a product that foreign competitors can bring into the country. For example, a quota on Japanese automobile imports limits the number of cars Japanese automakers can sell in the United States. This reduces competition in the domestic automobile market, which results in higher car prices, higher profits for domestic firms, and lower consumer surplus for domestic consumers. Domestic producers thus benefit at the expense of domestic consumers and foreign producers.

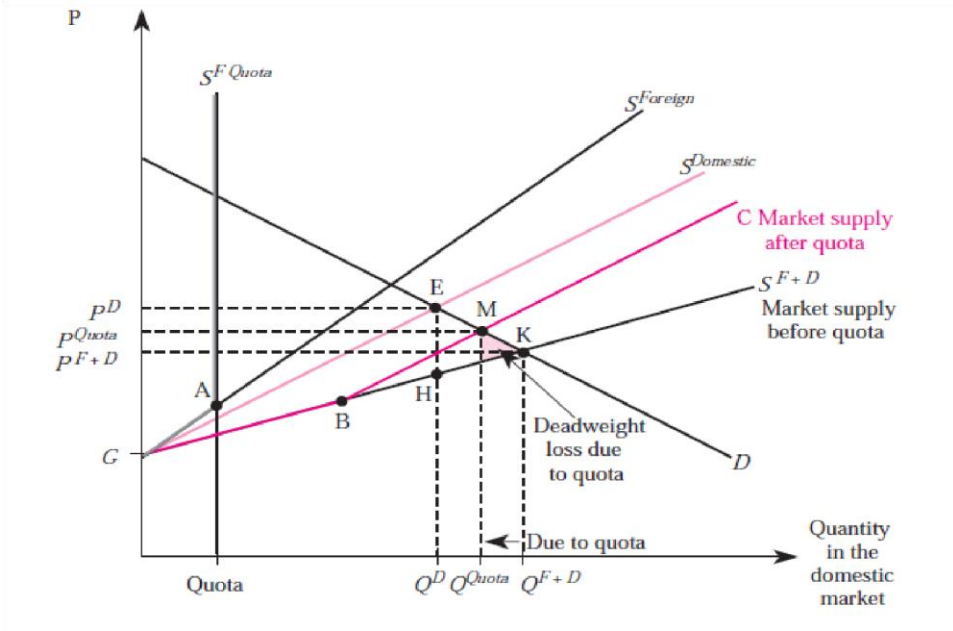


*Figure 6.5. The impact of a foreign quota on the domestic market*

To see why these results, occur, consider Figure 6.5, which shows the domestic market for a product. Before the imposition of a quota, the domestic demand curve is *D*, the supply curve for foreign producers is $S^{Foreign,}$ the supply curve for domestic producers is $S^{Domestic,}$ and the market supply curve—the horizontal summation of the foreign and domestic supply curves—is $S^{F+D}$. Equilibrium in the absence of a quota is at point K, where the equilibrium price is $P^{F+D}$ and the equilibrium quantity is $Q^{F+D}$.

Now suppose a quota is imposed on foreign producers that restricts them from selling more than the quota in the domestic market. Under the quota, foreign supply is $GASF^{Quota,}$ while the supply by domestic firms remains at $S^{Domestic.}$ Thus, market supply in the domestic market after the quota is *GBC*, resulting in an equilibrium at point M. The quota increases the price received by domestic producers to $P^{Quota,}$ and domestic firms now earn higher profits. The shaded triangle in Figure 6.5 represents the deadweight loss due to the quota. Total welfare declines as a result of the quota even though domestic producers earn higher profits. The reason for the decline in total welfare is that domestic consumers and foreign producers are harmed more than domestic producers gain from the quota. Domestic producers therefore have a strong incentive to lobby for quotas on foreign imports into their market.

*Demonstration problem*

Suppose the supply of a good by domestic firms is $Q^{SD} = 10 + 2P$ and the supply by foreign firms is $Q^{SF} = 10 + P$. The domestic demand for the product is given by $Q^d = 30 — P$.

1. In the absence of a quota, what is the total supply of the good?
2. What is the equilibrium price and quantity of the good?
3. Suppose a quota of 10 units is imposed. What is the total supply of the product?
4. Determine the equilibrium price in the domestic market under the quota of 10 units.

1. With a quota of 10 units, foreign firms will sell only 10 units in the domestic market. Thus, total supply is

$$Q^T = Q^{SD} + Q^{SF} = (10+2P) + 10 = 20+2P$$

2. Equilibrium is determined by equating total demand and total supply under a quota:

$30 — P = 20 + 2P$

Solving for $P$ yields the equilibrium price of $P$ = \$3.33. The quota increases the price of the good in the domestic market due to the reduction in foreign competition.

### 6.4.2. Tariffs

Tariffs, like quotas, are designed to limit foreign competition in the domestic market to benefit domestic producers. The benefits to domestic producers accrue at the expense of domestic consumers and foreign producers.

We will address two types of tariffs: lump-sum tariffs and excise or per-unit tariffs. A *lump-sum tariff* is a fixed fee that foreign firms must pay the domestic government to be able to sell in the domestic market. In contrast, a *per-unit* or *excise tariff* requires the importing firms to pay the domestic government a fee on each unit they bring into the country.

- **Lump-Sum Tariffs**

Figure 6.6 shows the marginal and average cost curves for an individual foreign firm before and after the imposition of a lump-sum tariff. The first thing to notice about the lump-sum tariff is that it does not affect the marginal cost curve. This is because the importer must pay the same amount of tariff regardless of how much of the product it brings into the country. Since the lump-sum tariff raises average costs from $AC^1$ to $AC^2$, an importer is unwilling to pay the tariff to enter the domestic market unless the price in the domestic market is at least $P^2$.

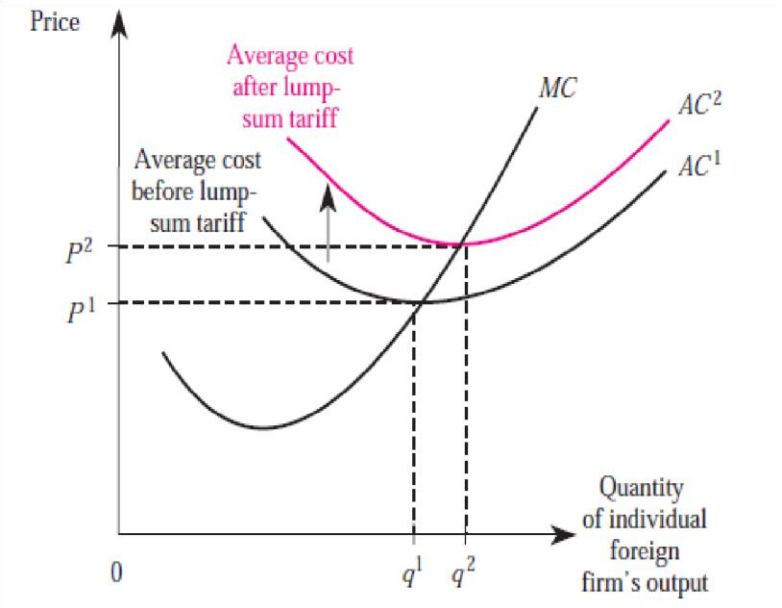


*Figure 6.6. Impact of a Lump-Sum Tariff on a Foreign Firm*

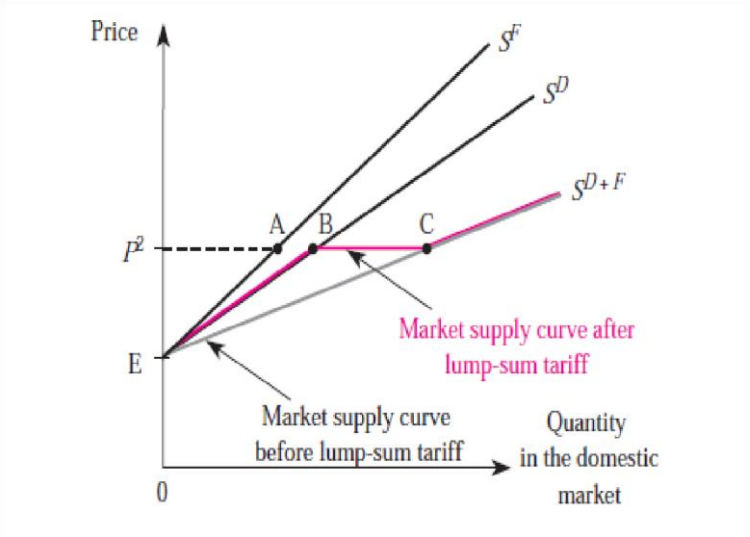


*Figure 6.7. Impact of a Lump-Sum Tariff on Market Supply*

Figure 6.7 shows the effect of a lump-sum tariff on the market. Before the tariff is imposed, the supply curve for foreign competitors is $ES^F$, that for domestic producers is $ES^D$, and the market supply curve, the summation of domestic and foreign supply, is $ES^{D+F}$. After the lump-sum tariff is imposed, the foreign supply curve becomes $AS^F$, because importers will not pay the lump-sum tariff to enter the domestic market unless the price is above $P^2$. Thus, the market supply curve in the presence of a lump-sum tariff is given by $EBCS^{D+F}$. The overall effect of this policy is to remove foreign competitors from the domestic market if the demand curve crosses the domestic supply curve at a price below $P^2$. A lump-sum tariff increases the profits of domestic producers if demand is low but has no effect on their profits if demand is high.

### Excise Tariffs

If an excise tariff is imposed on foreign producers instead of a lump-sum tariff, domestic producers benefit at all levels of demand. To see this, consider Figure 6.8, which shows the effect of an excise tariff. $S^F$ is the supply by foreign producers before the tariff, $S^D$ is the supply by domestic producers, and $ABS^{D+F}$ is the market supply curve before an excise tariff. Equilibrium in the absence of a tariff is at point H.

When a tariff of $T$ is imposed on each unit of the product, the marginal cost curve for foreign firms shifts up by the amount of the tariff, which in turn decreases the supply of all

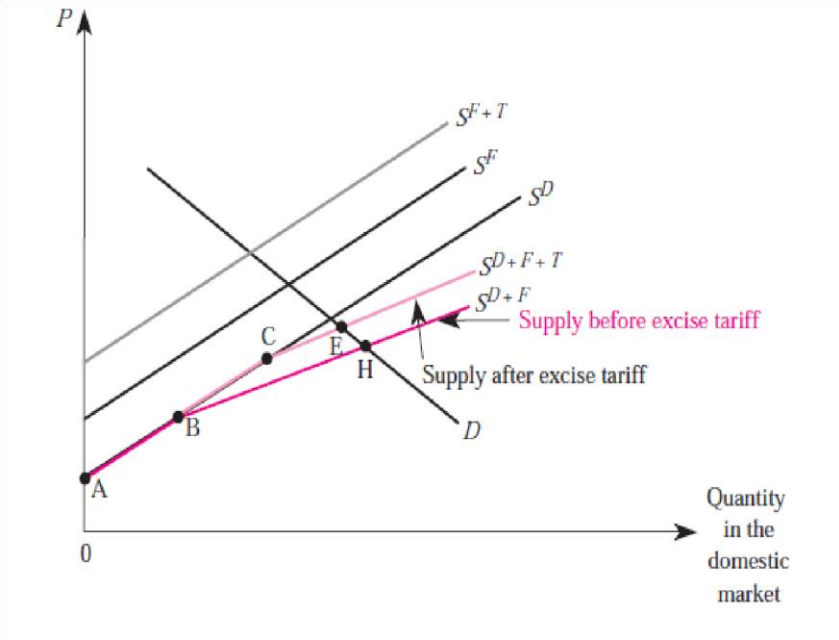


*Figure 6.8. Impact of an Excise Tariff on Market Supply*

foreign firms to $S^{F-T}$ in Figure 6.8. The market supply under a per-unit tariff is now $ACS^{D-F-T}$, and the resulting equilibrium is at point *E*. The tariff raises the price domestic consumers must pay for the product, which raises the profits of domestic firms at the expense of domestic consumers and foreign producers.

***Self Assessment Exercises***

1.You are an industry analyst that specializes in an industry where the market inverse demand is $P = 100 - 5Q$. The external marginal cost of producing the product is $^{MC}External = 10Q$, and the internal cost is $^{MC}Internal = 20Q$.

a) What is the socially efficient level of output?
b) Given these costs and market demand, how much output would a competitive industry produce?
c)Given these costs and market demand, how much output would a monopolist produce?
d)Discuss actions the government could take to induce firms in this industry to produce the socially efficient level of output.

1.As the manager of a monopoly, you face potential government regulation. Your inverse demand is $P = 25 - Q$, and your costs are $C(Q) = 5Q$.

a) Determine the monopoly price and output.
b) Determine the socially efficient price and output.
c)What is the maximum amount your firm should be willing to spend on lobbying efforts to prevent the price from being regulated at the socially optimal level?

2)Consider a competitive market served by many domestic and foreign firms. The domestic demand for these firms 'product is $Q^d = 500 - 1\ .5P$. The supply function of the domestic firms is $Q^{SD} = 50 + 0.5P$, while that of the foreign firms is $Q^{SF} = 250$.

a) Determine the equilibrium price and quantity under free trade.
b) Determine the equilibrium price and quantity when foreign firms are constrained by a 100-unit quota.
c)Are domestic consumers better or worse off as a result of the quota?
d)Are domestic producers better or worse off as a result of the quota?